A compact RGB-D map representation dedicated to autonomous navigation

Tawsif Ahmad Hussein Gokhool

UNIVERSITÉ DE NICE - SOPHIA ANTIPOLIS

ÉCOLE DOCTORALE STIC
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

T H È S E

pour l'obtention du grade de

Docteur en Sciences

de l'Université de Nice-Sophia Antipolis
Mention : AUTOMATIQUE ET TRAITEMENT DES SIGNAUX ET DES IMAGES

Présentée et soutenue par

Tawsif GOKHOOL

Cartographie dense basée sur une représentation compacte RGB-D dédiée à la navigation autonome

Thèse dirigée par Patrick RIVES

préparée à l'INRIA Sophia Antipolis, Équipe LAGADIC-SOP

soutenue le 5 Juin 2015

Jury:

Rapporteurs :	Pascal VASSEUR	-	Université de Rouen
	Frédéric CHAUSSE	-	Institut Pascal Clermont Ferrand
Directeur :	Patrick RIVES	-	INRIA Sophia Antipolis
Président :			
Examinateurs :	Noëla DESPRÉ	-	Airbus Defense and Space
	El Mustapha MOUADDIB	-	Université de Picardie, Amiens
	Cédric DEMONCEAUX	-	Université du Creusot
	Alessandro CORREA-VICTORINO	-	Université de Technologie de Compiègne

UNIVERSITY OF NICE - SOPHIA ANTIPOLIS

STIC DOCTORAL SCHOOL

INFORMATION AND COMMUNICATION TECHNOLOGIES AND SCIENCES

T H E S I S

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in **AUTOMATICS, SIGNAL AND IMAGE PROCESSING**
of the University of Nice - Sophia Antipolis

Defended by

Tawsif GOKHOOL

A Compact RGB-D Map Representation dedicated to Autonomous Navigation

Thesis Advisor: Patrick RIVES

prepared at INRIA Sophia Antipolis, LAGADIC-SOP Team

defended on 5^{th} of June, 2015

Jury:

Rapporteurs :	Pascal VASSEUR	-	University Rouen
	Frédéric CHAUSSE	-	Pascal Institute Clermont Ferrand
Directeur :	Patrick RIVES	-	INRIA Sophia Antipolis
Président :			
Examinateurs :	Noëla DESPRÉ	-	Airbus Defense and Space
	El Mustapha MOUADDIB	-	Universiy of Picardie, Amiens
	Cédric DEMONCEAUX	-	University of du Creusot
	Alessandro CORREA-VICTORINO	-	Compiègne Technological University

Acknowledgments

The work reported in this manuscript has been fully funded by Airbus Defence and Space group (ex Astrium/EADS) in consortium with INRIA Sophia Antipolis. It is indeed a privilege and a great achievement in itself for me to have been given the opportunity to be at the forefront of the research field at INRIA. The working ethics over here have instilled in me some important qualities such as independency, curiosity, drive and passion which I believe are important acquisitions to help building my career towards an established researcher. I take with me some wonderful memories of the time spent in the region of "la Côte d'Azure" whose breathtaking mountainous and coastal reliefs having helped me to overcome the home sickness of my Island Mauritius.

This thesis has been possible thanks to the trust bestowed upon me by my director, Patrick Rives. I cannot express my gratitude enough for his laudable support and for leading me till the end of this work. I would equally like to thank my "offline" mentor, Maxime Meilland in whom I found a great friend and a great guide. A special mention goes to Silvia Paris, Daniele Pucci, Glauco Scandaroli, Luca Marchetti and Alexandre Chapoulie for incubating me in their midst during my stay in France. My special thoughts go to all the members of the Lagadic and Hephaïstos team with whom I have shared some wonderful moments which shall be truly treasured. I thank Eduardo Fernández Moral, Renato Martins and Noëla Despré (Airbus Defense and Space) for collaborating with me. A big thank you to Romain Drouilly for listening to my qualms and for imbibing me the confidence of surmounting the slopes. Equally, a big thanks to Panagiotis Papadakis for always taking the time to review my work on demand. A huge thanks to Laurent Blanchet for providing the translated version of the introduction and the conclusion.

The Politecnico di Torino has been an important cornerstone in my academic life. My special accolades go to all the professors who were indulged in making the Master in Automatica and Control Technologies (MACT) so successful in the year 2011. The idea of diving into a phd was genuinely sparked on the bench of the Polito. My special thoughts go to MACT '11 Alumni for having made life so enjoyable in the beautiful city of Turin.

My international expedition began in the city state of Singapore in 2008 where I pursued my post graduate degree. My first experience away from home was fulfilled across South-East Asia where I was accustomed to an entirely different culture, a completely different walk of life. I have had the privilege to meet some true well wishers in Jawwad Akhtar, Yousouf Kamal, Danish Maqbool and Shehzaad Muhammad who pulled me along in their cart as we explored new horizons in the east.

I finally dedicate this thesis to my entire family based between Mauritius and France, who have been both financially and psychologically supportive during my entire academic journey till now. My parents have been a great source of inspiration to the way that I have led my life so far. I shall be truly indebted to them for all the sacrifices they have made for me. My all time buddy, Nandish Calchand has always been a source of light for me and has accompanied me in all aspect of decision making in my life. A huge shout out and a

big thank you mate for making me dream.

Lagadic Sop Team

Figure 1: *From left to right: Eduardo Fernández Moral, Renato Martins, Patrick Rives (Director), ME, Romain Drouilly*

Contents

1 Preamble **9**

1.1 Introduction . . . 9

1.2 This Thesis . . . 12

1.2.1 Manuscript organisation . . . 13

1.2.2 Publications . . . 15

2 State of the Art **17**

2.1 Introduction . . . 17

2.2 VO & SLAM . . . 17

2.2.1 Feature based vs Dense based . . . 19

2.2.2 Visual SLAM . . . 21

2.3 Map representation . . . 23

2.3.1 Occupancy grids . . . 23

2.3.2 Feature maps . . . 25

2.3.3 Topological maps . . . 26

2.3.4 Topometric maps . . . 27

2.3.5 Semantic Maps . . . 27

2.4 Towards Lifelong mapping . . . 28

2.4.1 Memory Models . . . 29

2.4.2 Graph pruning/ optimisation . . . 32

2.5 Conlusion . . . 34

3 Projective Geometry and Spherical Vision **37**

3.1 Introduction . . . 37

3.2 The camera model . . . 38

3.2.1 Intrinsic parameters . . . 38

3.2.2 Extrinsic parameters . . . 40

3.2.3 Projection equation . . . 41

3.3 Calibration . . . 42

3.3.1 Calibration with a planar pattern: . . . 42

3.3.2 Computation of Homography Transform: . . . 43
3.3.3 Computation of the Calibration Matrix: . . . 44
3.4 Stereo Calibration . . . 46
3.4.1 Epipolar geometry . . . 46
3.4.2 Image Rectification . . . 47
3.4.2.1 Calculation of matrix $\mathbf{R}'$ . . . 48
3.4.2.2 Point to Point correspondance . . . 48
3.4.2.3 Triangulation . . . 49
3.4.2.4 Pose recovery . . . 49
3.5 Spherical Perspective projection . . . 50
3.6 Image Warping: Novel View Synthesis . . . 51
3.6.1 A formal description . . . 51
3.7 Spherical Panorama . . . 52
3.7.1 Novel System Design and Calibration . . . 54
3.7.2 An innovative indoor Spherical RGBD sensor design . . . 57
3.8 Conclusion . . . 59

4 Spherical RGB-D Odometry 63
4.1 Introduction . . . 63
4.2 From 2D Optic Flow to 3D Scene Flow . . . 64
4.2.1 Direct Photometric Registration . . . 66
4.2.1.1 Optimisation Tools . . . 66
4.2.1.2 Efficient Second Order Minimization (ESM) . . . 68
4.2.1.3 Spherical Photometric cost function . . . 69
4.2.2 Rigid Body Motion . . . 71
4.2.3 Weighting Functions . . . 73
4.2.4 Information Selection . . . 74
4.2.5 Multi-Pyramid resolution . . . 76
4.3 Geometric constraint for motion estimation . . . 79
4.3.1 Direct Depth Map Alignment . . . 80
4.3.2 Point to plane registration . . . 83
4.4 Motion Tracking englobing *Photo* + *Geo* constraints . . . 84
4.4.1 Cost Function Formulation . . . 84

4.5 Keyframe-based Representation . . . 86
4.5.1 Median Absolute Deviation . . . 86
4.5.2 Differential Entropy . . . 87
4.6 Evaluation Metrics . . . 88
4.7 Results and Discussion . . . 90
4.7.1 Synthetic dataset: . . . 90
4.7.2 Inria Semir dataset . . . 92
4.7.3 Inria Kahn building dataset (ground floor) . . . 95
4.7.3.1 Metric loop closure . . . 97
4.7.4 Results with Garbejaire Dataset . . . 100
4.8 Conclusion . . . 102

5 Towards Accurate and Consistent Dense 3D Mapping 103
5.1 Introduction . . . 103
5.2 Methodology . . . 103
5.3 A first approach to depth map fusion . . . 104
5.3.1 Depth inconsistency detection . . . 104
5.3.2 Inverse Warping and Depth Map Fusion . . . 106
5.3.3 Results . . . 108
5.3.4 Pose graph optimisation . . . 110
5.4 An improved approach to environment modelling . . . 112
5.4.1 Error modelling and propagation . . . 114
5.4.2 Homogeneous Vector Uncertainty . . . 115
5.4.3 Warped Sphere Uncertainty . . . 115
5.4.3.1 Indoor spherical sensor model . . . 117
5.4.4 Probabilistic data association . . . 118
5.4.5 Features/landmarks visibility scenario . . . 118
5.4.6 Formulation . . . 119
5.4.7 Dynamic points filtering . . . 122
5.4.8 Application to Saliency map . . . 124
5.4.9 Results . . . 125
5.4.10 Discussion . . . 128
5.5 Conclusion . . . 128

6 Conclusion and Perspectives **131**

6.1 Conclusion 131

6.2 Perspectives 133

Bibliography **141**

List of Figures

1 Team Presentation ii

2 European Commission projects 3

1.1 European Commission projects 11

2.1 Map representations 23

2.2 Pose/feature graph 25

2.3 Atkinson and Shiffrin human memory model 28

2.4 Long term/ Short term memory model 29

2.5 FSH memory model 31

3.1 Perspective projection 39

3.2 Transformation from image to pixel frame 40

3.3 Transformation from world to camera frame 41

3.4 Calibration with checkerboard pattern 43

3.5 Epipolar Geometry 47

3.6 Image rectification 47

3.7 Spherical perspective projection model 51

3.8 Camera under motion 52

3.9 Acquisition platform with multicamera system 53

3.10 Perspective Image transformation on a sphere 54

3.11 Novel synthesised spherical panoramic image 54

3.12 Spherical RGBD outdoor acquisition system 55

3.13 Augmented RGBD sphere 55

3.14 Spherical Triangulation 56

3.15 Multi RGBD indoor acquisition system 57

3.16 Spherical RGBD construction 57

3.17 spherical panoramic views from indoor multi camera rig 58

3.18 Calibration of non overlapping cameras 60

3.19 Calibration of non overlapping cameras 60

4.1 Scene flow to optic flow modelling 64

4.4 Application of saliency map . . . 76
4.2 Jacobian decomposition . . . 77
4.3 saliency table . . . 77
4.5 Multi pyramid resolution . . . 79
4.6 Depth map warping . . . 80
4.7 Principle of ICP . . . 83
4.8 Keyframe-based representation . . . 87
4.9 Illustration of Entropy ratio . . . 88
4.10 Synthetic dataset . . . 91
4.11 Trajectory comparison . . . 92
4.12 Performance comparison between cost functions . . . 93
4.13 Snapshots of Inria Semir dataset . . . 94
4.14 Reconstruction using Semir dataset . . . 94
4.15 Performance comparison between cost functions . . . 95
4.17 Reconstruction using Inria Kahn building dataset . . . 95
4.16 Comparison between Keyframe criteria . . . 96
4.18 Snapshots of Inria Kahn building dataset . . . 96
4.19 Pose graph correction . . . 97
4.20 Illustration of loop closure between two nodes . . . 99
4.21 Rotation estimation using an SSD cost function . . . 99
4.22 Reconstruction comparison with metric loop closure . . . 100
4.23 Trajectory reconstruction with Garbejaire dataset . . . 100
4.24 Full trajectory reconstruction . . . 101

5.1 Page-Hinckley Test: events . . . 105
5.2 Page-Hinckley Test: no events . . . 106
5.3 Inverse warping on reference frame . . . 107
5.4 Pipeline using P-H test . . . 108
5.5 Reconstruction comparison with Inria Kahn dataset using 2 with data fusion 108
5.6 Evualuation of Kahn0 dataset . . . 109
5.7 Trajectories with pose graph optimisation . . . 111
5.8 Results with pose graph optimisation . . . 113
5.9 Pipeline using P-H test . . . 114

5.10 probabilistic data association 119
5.11 Node comparison pre and post filtering 121
5.12 Filtered depth map evaluation 121
5.13 Sphere segmentation using SLIC 122
5.14 Saliency map skimming 124
5.15 Trajectory comparison with and wihout fusion using error model 125
5.16 Reconstruction comparison with Kahn dataset 126
5.17 Evualuation of Kahn0 dataset 126
5.18 Comparison between vision and laser maps 127

Préambule

Introduction générale

La manière dont les Hommes et les animaux interagissent avec la nature a toujours été fascinante. Cette marche, ce vol ou encore cette navigation dans leur environnement, ou autant de prouesses réalisées sans y réfléchir par le biais des capacités sensorielles respectives. Cette aisance inspire la communauté de la recherche en robotique mobile à reproduire de telles capacités au travers d'une certaine forme d'intelligence artificielle. Cette tendance du domaine des robots mobile a permis d'étendre le domaine d'interaction de l'Homme, là où l'interaction directe pour des tâches techniquement éprouvantes est considérée trop risquée ou hors de notre portée physique. Cette extension inclue des tâches aussi diverses que l'exploration de mines abandonnées ou de l'environnement martien, avec les « Mars Rovers » en 2004, jusqu'aux interventions à risques de la centrale nucléaire de Fukushima en 2011 ; toutes ces éminentes missions mettant en exergue les progrès du domaine de la robotique mobile.

Tout robot autonome dont la tâche principale est la navigation est confronté à deux difficultés principales, à savoir la découverte de l'environnement et sa propre localisation. Pour accomplir cette seconde tâche à partir des capteurs équipant ce robot, celui-ci nécessite la connaissance d'une carte de l'environnement. Pour découvrir ou redécouvrir l'environnement, et établir ou étendre la carte correspondante, une position précise sur la carte est nécessaire. Depuis deux décades, ces deux problèmes sont devenus la pierre angulaire d'un domaine de recherche connu comme « Simultaneous Localisation and Mapping (SLAM) », soit Cartographie et Positionnement Simultanés. Une carte permet de plus de réaliser des tâches planifiées, en fournissant au robot les informations nécessaires lui permettant, à partir d'une position originelle A, de se déplacer jusqu'à une certaine destination B.

Des solutions commerciales de véhicules autonomes sont produites spécifiquement pour la tâche donnée. L'aspirateur Roomba de la société iRobot est équipé d'un ensemble de capteurs basiques adaptés à certaines fonctions spécifiques, comme l'évitement d'obstacle, la détection de la présence de saletés au sol, ou encore de capteurs d'inclinaison pour éviter de tomber dans d'éventuels escaliers. Sur les quais industriels, des robots autonomes guidés sont devenus des composants clés des opérations de chargement et déchargement de navires [Durrant-Whyte 1996], induisant une meilleure gestion du trafic sur les quais, une productivité améliorée et des coûts opérationnels réduits. Les opportunités présentées par les véhicules sans conducteurs n'a pas plus laissé l'industrie minière indifférente, avec l'émergence d'une technologie baptisée « Autonomous Haulage System (AHS) » par Rio Tinto (une des principales sociétés du secteur) [AHS 2014]. Ces camions sans composante humaine sont particulièrement adaptés aux conditions difficiles des procédures minières, de part leur activité 24/7 ininterrompue. Les bénéfices rapportés sont colossaux ; de l'ordre de 15–20% d'augmentation de la production, de 10–15% de réduction de la consomma-

tion de carburant, et encore d'un gain de 8% en coût de maintenance. D'un point de vue des ressources humaines, cette technologie a apporté une modification fondamentale des conditions de travail en déplaçant les travailleurs de zones potentiellement dangereuses vers d'autres postes aux risques réduits.

Toutes ces applications qui viennent d'être citées, dédiées à leurs tâches respectives, ne sont fonctionnelles que dans leurs environnements contrôlés propres ; et certaines, particulièrement dans les domaines d'opération industriels, sont de plus dépendantes de l'infrastructure, avec des espaces de travail hautement contraints et modulaires. Il n'existe malheureusement pas d'application unique disponible pour toute tâche.

La difficulté principale des robots autonomes est sans aucun doute d'opérer au sein d'applications pratiques du monde réel. Il n'est par exemple pas possible de structurer l'environnement sous-marin profond par des infrastructures dédiées dans le but d'explorer la vie sous-marine. Il n'est guère plus envisageable d'un point de vue économique de placer des marqueurs sur l'ensemble de nos réseaux routiers pour aider à la conduite autonome. Aussi, la navigation autonome et la cartographie doivent pouvoir être effectuées à partir des seules informations sur l'environnement perçues par le robot. Dans cette optique, le Mobile Robotics Group de l'université d'Oxford [MRG] a établit son motto comme « l'extension à grande échelle de la portée de la navigation autonome, sans requérir de coûteuses, gênantes, et incongrues modifications de l'environnement ». Ils développent au sein d'un consortium réunissant aussi le fabricant de voitures Nissan une plateforme autonome équipée de technologies de laser et de vision. L'idée est alors de construire des cartes incluant des informations de situation par rapport à l'environnement statique et dynamique. Cette information sémantique est extraite de l'environnement statique que forment le marquage routier et les feux de signalisation, les informations de direction, ou encore les trottoirs. Ces informations nécessitent toutefois d'être mises à jour à cause des perpétuels changements autour de nous. Pour sa part, l'environnement dynamique consiste aux entités mobiles ou stationnaires telles que les voitures, vélos, piétons, et les autres obstacles. À partir d'une prédiction de comportement, le profil de conduite peut alors être établi en conséquence.

Le Grand Challenge lancé par la « Pentagon's Defense and Advanced Research Projects Agency » (DARPA) a été le théâtre du franchissement d'un important jalon pour les technologies des véhicules autonomes terrestres (Autonomous Ground Vehicules, AGVs). Lors de ce défi les participants devaient produire un AGV capable de concourir dans une course de 175 miles (282 kms) au travers du désert de Mojave (sud-ouest des États-Unis), et ce dans un temps imparti de 10 heures. Les AGVs participants ont été poussés dans leurs retranchements sur tous types de surfaces : des chemins boueux aux cols montagneux dentelés, en passant par l'éreintant sol désertique. Le groupe de Sebastian Thrun de l'université de Standford a été le grand gagnant de cette compétition à l'aide du robot Stanley. Pour capitaliser cette participation, Professeur Thrun s'est épaulé de Google pour leur ambitieux projet de véhicules robotisés hybrides, dans un consortium réunissant de plus Toyota. En Mai 2011, ce consortium comptabilisait [Gca 2011] 140K miles (225000 kms) de kilométrage cumulé de leurs six Toyota Prius autonomes et unique Audi TT parcourus sur les

routes de Californie, dont plus de 1000 miles (1500 kms) en complète autonomie. La technologie utilisée est perçue comme améliorant l'efficacité énergétique tout en réduisant les accidents et morts sur les routes.

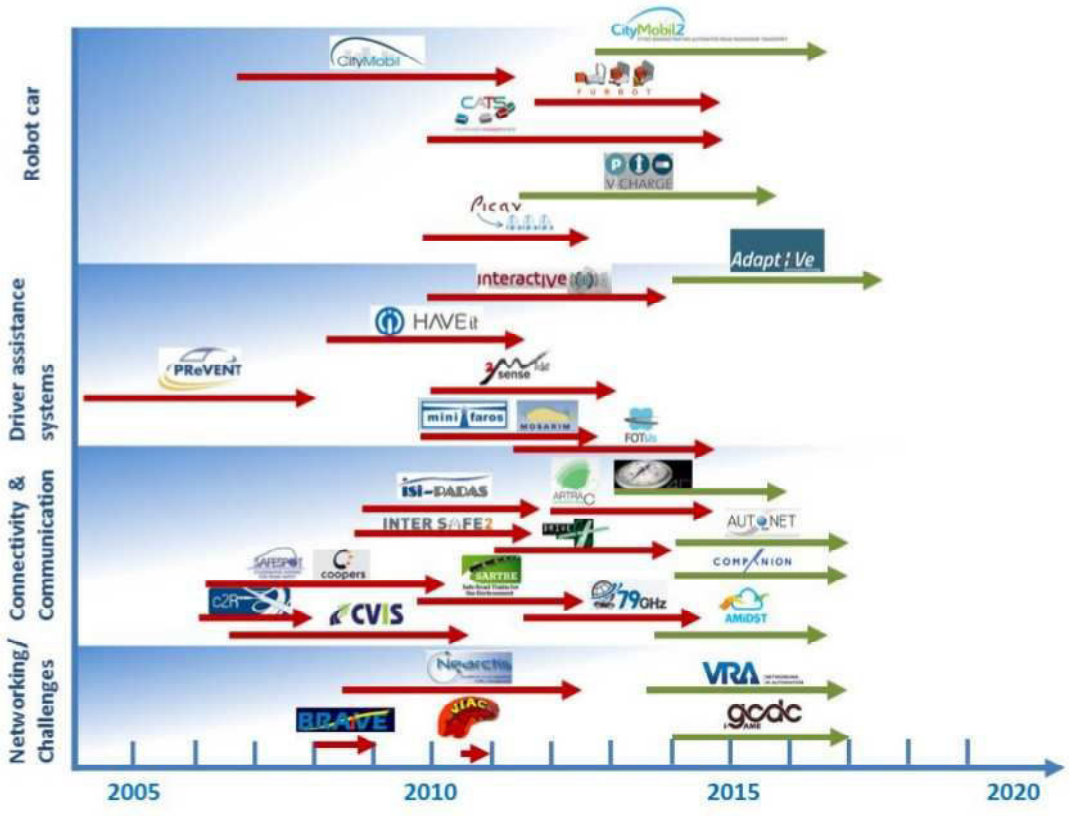

FIG. 2: *Vue d'ensemble des projets financés destinés au développement de la navigation autonome. Cette étude a été realisée pendant les dix dernieres années. Les flèches en rouges représentent des projets complétés et celles en vertes , les projets en cours (source : The European Technology Platform on Smart Systems Integration- EPoSS [EPo])*

Du côté Européen, de nombreux projets ont vu le jour dans des applications de voitures intelligentes et d'urbanisme. Un de ces projets des plus avancés est CityMobil2, successeur de CityMobil. Ce projet implique un système de transport routier public automatisé local appelé Cybernetic Transport System (CTS), qui agirait à la demande sur le même principe qu'un ascenseur. Son but serait de compléter les transports publics existants pour les cas de faible influence ou dessertes éloignées, offrant un service efficace et plaisant aux usagers routiers. Ces cyber-voitures pourraient alors rouler dans des zones particulières telles que les zones piétonnes, les parkings et sites privés tels que les hôpitaux, ou encore les voies de bus libres. Une première phase de test a été menée dans la ville de La Rochelle en France.

Le projet V-Charge [VCh] quant à lui, devrait apporter des solutions aux évolutions supposées des transports publics et personnels des années à venir. Leurs infrastructures incluent une aide à la conduite dans les environnements urbains, ou des options telles que la

conduite autonome dans certaines zones (stationnement automatisé, parc-relai). Une plateforme (semi-) robotique utilisant des capteurs ultrasoniques, de signal GPS, et caméra, a été conçue au sein de ce projet, comme un élément de cette contribution majeure. La figure 1.1 présente une liste exhaustive des projets financés par la commission européenne depuis 2005, terminés ou en cours.

Bien que le progrès technologique est permanent, avec plusieurs des projets prêt à être commercialisés, le plus gros obstacle à ce point de développement en devient le cadre législatif. Les incroyables avancées technologiques menacent de dépasser les lois existantes sur la mobilité et le transport, certaines desquelles datent de l'époque des carrioles à chevaux.Ainsi, pour que la conduite autonome prenne son essor comme une incontournable réalité dans un futur proche, les décideurs doivent rapidement réagir pour anticiper un fonctionnement cohérent de ces cyber-voitures.

Inscription de cette thèse dans le cadre présenté

Notre attention se concentre sur l'élaboration de cartes topographiques égo-centrées représentées par un graphe d'images-clés qui peuvent ensuite être utilisées efficacement par des agents autonomes. Ces images-clés constituant les nœuds de cette arborescence combinent une image sphérique et une carte de profondeur (couple que l'on appelle sphère de vision augmentée), synthétisant l'information collectée sur l'environnement immédiat par un système de capteurs embarqués. La représentation de l'environnement global est obtenue par un ensemble de sphères de vision augmentées, apportant la couverture nécessaire de la zone opérationnelle. Un graphe de « pose » liant ces sphères les unes aux autres dans un espace de dimension six définit le domaine potentiellement exploitable pour la navigation en temps réel. Nous proposons dans le cadre de cette thèse une approche de la représentation basée sur les cartes en considérant les points suivants :

- L'application robuste de l'odométrie visuelle, tirant le meilleur parti des données de photométrie et de géométrie disponibles par notre base de données des sphères de vision augmentée
- La détermination de la quantité et du placement optimal de ces sphères augmentées pour décrire complètement un certain environnement
- La modélisation des erreurs de mesure et la mise à jour des données compactes des sphères augmentées
- La représentation compacte de l'information contenue dans les sphères augmentées pour s'assurer de la robustesse, la précision et la stabilité le long d'une trajectoire, en exploitant les cartes d'intérêt

Cette recherche met à profit et étend les résultats du projet à succès CityVip présentés dans [Meilland *et al.* 2010, Meilland *et al.* 2011a, Meilland *et al.* 2011b], de part plusieurs aspects. Dans l'optique d'améliorer l'odométrie visuelle, une fonction de coût est introduite, combinant autant les contraintes géométriques que photométriques s'appliquant sur

la capture directe de l'image pour l'estimation de la pose. De plus le problème de l'optimisation du graphe de pose est abordé par laquelle à partir d'une base de sphères visuelles augmentées, une trajectoire explorée aux données redondantes est extraite pour former un graphe de pose squelettique épuré.

Cartographier sous l'hypothèse d'un environnement statique est sous un certain aspect sans intérêt, puisque l'environnement dans lequel le robot évolue nominalement est dynamique et évolue de manière imprédictible. Bien que certaines parties du milieu sont statiques en considérant un intervalle de temps court, d'autres sont susceptibles de changer abruptement selon les activités se déroulant tout autour du robot, comme des piétons en n'importe quel point voisin et les voitures sur la route. Cet aspect est couvert dans notre travail par l'introduction du concept des entités dynamiques qui évoluent selon une trajectoire. En plus du graphe initial des sphères de vision augmentée, comprenant les cartes photométrique, géométrique, et saillance, deux composants supplémentaires sont maintenant liés aux contenus d'information sur l'environnement de notre graphe ; à savoir la carte d'incertitude et la carte de stabilité.

Organisation du document

Le chapitre 2 établit l'état de l'art des systèmes de SLAM uniquement basés sur la vision. Les écueils d'autres technologies de capteurs comme le GPS, les encodeurs des axes des roues, ou encore des scanners laser, ont poussé les chercheurs à exploiter l'horizon des possibilités offertes par les contenus riches des images. L'historique du développement de l'odométrie visuelle (VO) montre que les efforts initiaux portaient tout particulièrement sur la construction de robots mobiles robustes destinés à l'exploration planétaire. Ensuite vinrent les applications de véhicules terrestres, aériens, et sous-marins intelligents, diversifiant ces techniques d'odométrie visuelle. Un autre domaine très prometteur requérant la VO est celui de la réalité augmentée. Tout comme d'autre techniques d'odométrie, la VO est sujette au problème de dérives. Dans la littérature, la VO est décrite par deux approches différentes : une approche par entité d'intérêt, et une approche par densité. Les avantages et inconvénients des deux techniques sont soulignés ; des exemples tirés de la littérature sont détaillés pour mettre en exergue la place de la VO dans le cadre du SLAM. Dans un but de positionnement, un robot requiert normalement une carte de son environnement perçu. Cette carte peut être établie au fil de la phase d'exploration, ou séparément au cours d'une phase d'apprentissage initiale d'exploration pure. Dans ce contexte, la cartographie est constituée de plusieurs couches : topologie métrique, sémantique. Les avantages et inconvénients de ces couches sont ensuite élaborés plus avant. Enfin, un mot doit être dit sur le domaine naissant des cartes permanentes. Cette approche cherche à équiper des véhicules intelligents de nouveaux outils pour appréhender les défis de l'exploration à grande échelle s'appuyant sur des ressources telles que la capacité de stockage, de puissance de calcul, ou de navigation complètement autonome et non-supervisée, limitées. Malgré ces limitations, le robot doit avoir la capacité de créer une carte stable, qui peut alors être réutilisée autant de fois qu'elle est nécessaire sur de longues périodes de temps.

Le chapitre 3 introduit le lecteur au monde de la vision 3D dans la première partie. Dans le but d'illustrer le processus complet de formation de l'image sur l'optique de la caméra, une application directe de calibration de caméra est discutée. Pour ensuite extrapoler les informations 3D comme perçues par la vision humaine à partir des images 2D, le concept de vision stéréo est mis en avant en décrivant l'ensemble des phases intermédiaires jusqu'à l'obtention des informations de profondeur. Cette synthèse concise de géométrie projective et de vision stéréo permet d'établir certains concepts de base qui seront utilisés par la suite dans la deuxième partie du chapitre où la vision sphérique est introduite. On établit l'intérêt de la représentation par vision sphérique et ses multiples avantages, c'est à dire un modèle enrichi compact de l'environnement, avec une représentation par champ de vision (Field of View, FOV) omnidirectionnel à 360^0 et invariant par rotation, ce qui rend cette représentation imperméable aux différentes configurations de capteurs. De plus, couvrir l'environnement exploré sur le plus grand intervalle possible et d'orientations possibles, conduit à une meilleure localisation. Ces avantages ont motivé les chercheurs à élaborer des technologies innovantes telles que les caméras catadioptriques, ou encore mieux, les systèmes de caméras multi-trames. Tandis que ces premiers systèmes permettent d'obtenir immédiatement des images sphériques, l'obtention d'un panorama sphérique avec les seconds n'est pas direct. Cette difficulté requiert des opérations intermédiaires sur les images, telles que déformation, fusion, et composition. Dans ce chapitre, une vue d'ensemble de ces techniques est donnée, avec des applications aux milieux en intérieur et en extérieur des systèmes multi-capteurs développés dans le cadre des activités de recherche de l'équipe. Pour chaque système, la technique de calibration propre est surimposée, produisant les matrices extrinsèques des systèmes multi-caméra. Cette étape est essentielle pour la production de panorama sphériques virtuels rendant la vue équivalente à ce qui devrait être vu.

Le chapitre 4 discute de l'odométrie RGB-D sphérique. On commencera par une description compréhensive du modèle du flux optique, extrapolé en un flux scénique 3D pour l'application directe à l'estimation de mouvement. Ce concept des plus importants structure de la technique Lucas-Kanade traitant de la détection directe basé sur les images, laquelle calcule un mouvement paramétré relativement inconnu entre deux images, pour un déplacement relativement faible entre les deux images. Un aperçu de la fonction de coût photométrique est produit en appliquant celle-ci à notre ensemble de sphères visuelles augmentées, lesquelles consistent en images sphériques RGB et cartes de profondeur et d'intérêt correspondantes. Les techniques basées sur l'intensité montrent leurs limitations dans des environnements soit mal éclairés, soit sujets à de grandes variations de luminosité. Pour compenser les manquements de ces méthodes, une seconde fonction de coût est introduite, prenant en compte explicitement le contenu géométrique des cartes de profondeur par l'implémentation d'une technique itérative du point le plus proche (Iterative Closest Point, ICP) d'un plan, dont l'inspiration provient de la littérature. La minimisation hybride de ces deux fonctions de coût conduit à une formulation améliorée incluant autant des informations géométriques que photométriques. L'environnement global est représenté par

un graphe de pose constitué de nœuds et d'arêtes, et établit par la VO sphérique ; chaque nœud est représenté par une image-clé. La sélection de ces images-clés est primordiale dans une telle représentation, puisqu'une sélection avisée résulte en plusieurs avantages. Tout d'abord, le redondance des données est minimisée, minimisation impliquant une représentation plus compacte et clairsemée. Ces deux propriétés sont des plus recherchées dans le cadre de l'exploration d'un environnement de grande échelle et avec une capacité limitée. Ensuite, l'utilisation d'un nombre réduit d'images-clés facilite la réduction de l'intégration des erreurs de dérive inhérentes à l'odométrie d'image à image-clé. Dans la même optique, deux autres critères notables sont considérés, la Déviation Absolue Médiane (Median Absolute Deviation, MAD) et un second utilisant la différence d'entropie, laquelle est une image de de l'incertitude de la pose. Pour valider les différents aspects discutés dans le chapitre, une partie « résultats » dédiée détaille quatre cas : deux ensembles de données simulées et deux ensembles de données réelles. Les avantages et limites de l'algorithme sont ensuite discutées avant de conclure. L'évaluation expérimentale de ce chapitre a été en partie présentée dans une conférence internationale [Gokhool *et al.* 2014] et une seconde, nationale [Rives *et al.* 2014].

Le chapitre 5 aborde les diverses limites précédemment établies ; notamment les problèmes de dérive ou de perte de l'odométrie qui casse alors la cohérence globale de la représentation par graphe de pose/d'intérêt. Une première approche cherche à améliorer la carte de profondeur, au profil très bruité résultant du capteur. Pour pouvoir fusionner les mesures de la carte de profondeur, celles-ci doivent être représentées dans un référentiel commun. Les observations obtenues le long d'une trajectoire sont obtenues dans ce référentiel particulier par transformation inverse. Pour détecter les incohérences dues à des problèmes d'occultation ou de bruit entre les cartes de profondeur mesurée et observée, un test statistique est ajouté à l'implémentation. Ce test consiste principalement à la détection de la dérive sur la base des séries de temps moyennes d'un signal discret. L'intérêt limité de cette méthode lors de sa première évaluation a conduit à considérer une seconde, prenant en compte les erreurs aléatoires et dynamiques des capteurs. Un modèle d'incertitudes bien meilleur est formulé en prenant en compte autant la conception de notre représentation sphérique que les incertitudes de l'opération de transformation. Une technique d'association de données probabiliste est conçue pour identifier les points aberrants dus aux phénomènes de bruit, occultations, et violation de l'espace libre. Les valeurs observées correspondant aux mesures sont fusionnées pour améliorer le contenu géométrique et photométrique des sphères augmentées. Le cas des points dynamiques est de plus traité, conduisant ainsi à création d'une nouvelle entité dans notre ensemble de de sphères augmentées, la carte de stabilité. Cet attribut supplémentaire est utilisé pour mettre à jour la carte d'intérêt. Une partie expérimentale évalue cette seconde approche, avant de conclure. La formulation et l'évaluation expérimentale correspondante ont été en partie présentée dans une conférence internationale [Gokhool *et al.* 2015].

Le chapitre 6 résume l'évaluation de ce travail et des perspectives sont établies et discutées pour permettre de d'amener ce projet de cartographie encore plus loin en utilisant de plus le cadre VSLAM sous-jacent.

Contributions

Internationale :

- T. Gokhool, R. Martins, P. Rives and N. Despré. *A Compact Spherical RGBD Keyframe-based Representation*. In International Conference on Robotics and Automation, (ICRA), Seattle, Etats Unis, May 2015.
- T. Gokhool, M. Meilland, P. Rives and E. Fernàndez-Moral. *A Dense map Building Approach from Spherical RGBD Images*. In International Conference on Computer Vision Theory and Applications, (VISAPP), Lisbon, Portugal, January 2014.

Nationale :

- P. Rives, R. Drouilly and T. Gokhool. *Représentation orientée navigation d'environnements à grande échelle.* Reconnaissance de Formes et Intelligence Artificielle, RFIA 2014, France, June 2014.

CHAPTER 1

Preamble

1.1 Introduction

We have always been captivated in the way humans and animals interact with nature. The way they effortlessly walk, fly or navigate in the environment using their sensing prowess. We have indeed been inspired to replicate the sort of capabilities in robots by fitting them with the kind of artificial intelligence. Having said so, the field of mobile robotics has extended our reach to areas where human investigation is considered too risky or beyond our physical means as the tasks presented are too technically challenging. Applications ranging from exploration of abandoned mines, to planetary missions of Mars Rovers in 2004, to hazardous interventions in the Fukushima nuclear plant in 2011, all these valuable missions epitomises the progress made in the area of mobile robotics.

The core challenge for any autonomous robot whose main task is navigation is composed of two characteristic problems; environment mapping and localisation. To be able to locate itself by using onboard sensors, the robot requires a map. Parallely, to be able to upgrade or to extend the map, precise location on the map is required. Since two decades,these two problems have become the cornerstone of a research field known as *Simultaneous Localisation and Mapping* (SLAM). A map further bolsters trajectory planning tasks by providing the robot with the required information to drive from source place A to destination B.

Commercial autonomous vehicle solutions are tailormade accordingly to the task at hand. The Roomba vacuum cleaner of iRobot is equipped with a set of basic sensors curtailed for specific functions, such as obstacle avoidance, detection of dirts on the floor, steep sensing to prevent falling off down stairs. In cargo handling terminals, autonomous guided vehicles have been a key component in ship loading and unloading [Durrant-Whyte 1996], leading to more efficient traffic management in cargo terminals, increased productivity and reduced operating costs. The wonders of driverless vehicles have not left the mining industry indifferent with the emergence of a technology baptised as Autonomous Haulage System (AHS) by Rio Tinto (a leading company in the business) [AHS 2014]. These unmanned trucks adapt well to the rigorous mining procedures, providing an incessant $24/7$ service round the clock. Reported benefits are huge; $15-20$ percent increase in output, $10-15$ percent decrease in fuel consumption, 8 percent gain in maintenance cost. On the workforce plan, this has led to a shift in manpower requirements by removing workers from potentially hazardous environments to new employment opportunities with reduced

labour intensive operations.

The above-mentioned task specific applications are functional in controlled environments and some of these solutions, especially those operating in the industry are infrastructure dependent with constrained and highly manageable workspaces. Unfortunately, there does not exist a single off-shelf application for all applications. Undoubtedly, the most challenging task for autonomous robots is to operate in practical real world applications. It is not feasible to fix dedicated infrastructures for deep sub-sea navigation for the exploration of marine life for example. It is economically not viable to place artificial markers along our road networks to facilitate autonomous driving. Therefore, autonomous navigation and mapping should be done by relying solely on the environment that the robot perceives. In this context, the Mobile Robotics Group at Oxford University [MRG] aims at *"extending the reach of autonomous navigation to vast scales without the need of expensive, awkward and inconvenient modification of the environment"*. In consortium with Nissan car manufacturer, an autonomous platform is developed with laser and vision technology. The idea is to build maps incorporating situational awareness with respect to static and dynamic environments. Semantic information is extracted from static environments such as road marking, traffic lights, lane information, curbs. These information however, needs updating over time due to the constant nature of changes around us. Dynamic information integrates moving or stationary entities such as cars, bicycles, pedestrians, obstacles. Based on their predictive behaviour, the driving profile can be moulded accordingly.

The Grand Challenge launched by Pentagon's Defense and Advanced Research Projects Agency (DARPA) saw a great milestone laid out in autonomous ground vehicles (AGV)s. Participants were required to come up with AGVs to compete in a race track of 175 miles in the Mojave desert (south west US) within a limited time frame of 10 hours. The depth and breadth of AGVs were severely put to test on all types of surfaces; from dirt roads to rugged mountaineous passages, not to forget the sluggish nature of the desert. Sebastian Thrun's group at Stanford University emerged out as big winners for this competitions thanks to the robot Stanley. To take forward this fruitful experience, Professor Thrun recently joined hands with Google for their ambitious robotic hybrid vehicle project (in consortium with Toyota). It was reported that, as at May 2011 [Gca 2011] the registered mileage of Google's six fleet of six autonomous Toyota Priuses and one Audi TT stood at 140K miles on the roads of California, with more than 1000 miles performed on fully autonomous mode. The implemented technology is hyped to increase energy efficiency whilst reducing road injuries and deaths rates in the future.

On the European level, many projects have seen the lights too with applications in cybercars and urban technology. CityMobil2, an offspring of CityMobil is now well under way. In this project, a local public automated road transportation system, baptised as CTS-Cybernetic Transport System, works on demand with the same principle of operation of an elevator. The purpose is to complement public mass transportation systems when demand is low or pick-up points are far apart providing a more effective and pleasant service to road users. These cybercars will be allowed to run in dedicated areas such as pedestrian

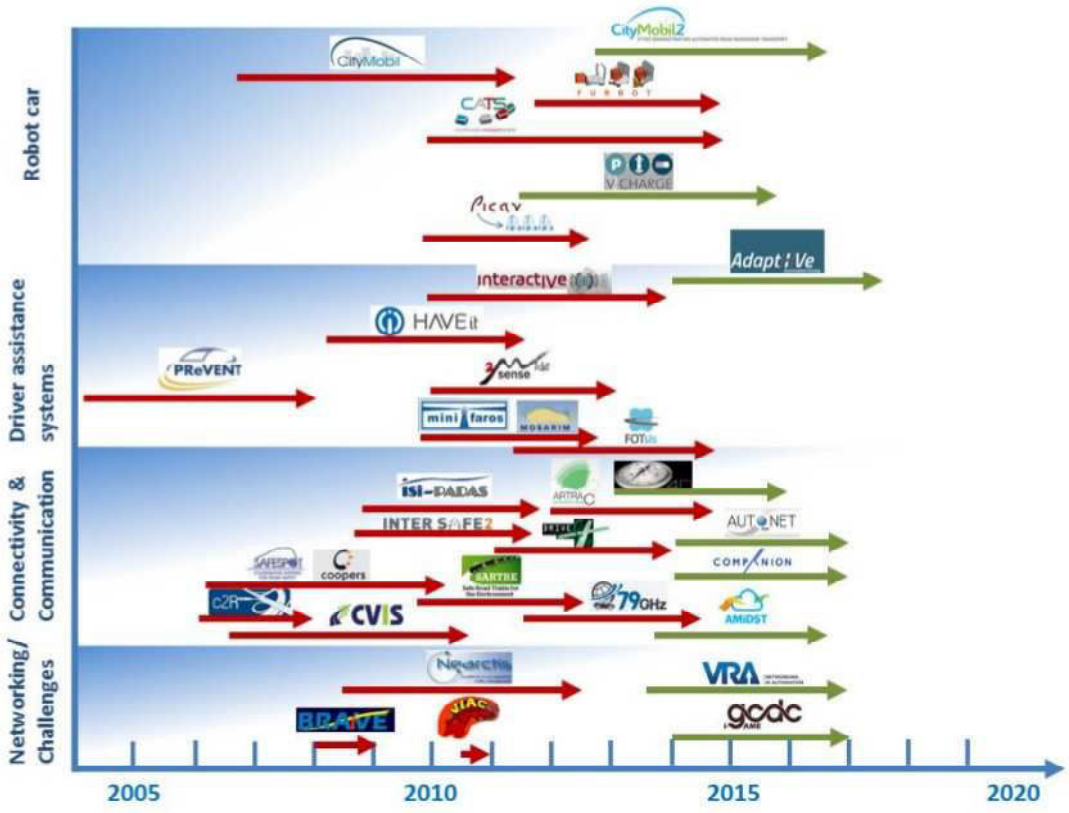

Figure 1.1: *An overview of funded projects that support the development of automated driving. The analysis has been done for the period of the last ten years. Red arrows correspond to completed projects and green arrows relate to ongoing projects (source: The European Technology Platform on Smart Systems Integration- EPoSS [EPo])*

areas, car parks, private sites such as hospitals or dedicated bus lanes when unoccupied. An initial testing phase has been conducted in the city of La Rochelle in France.

The V-Charge project [VCh], too is envisioned on providing solutions for anticipated changes in public and individual transportation in the years to come. The facilities include: advanced driver support in urban environments, options such as autonomous driving in designated areas (*e.g.* valet parking, park and ride) will also be offered. As part of this key contribution, a (semi) robotic platform has been conceived by using cost GPS, camera and ultrasonic sensors. Figure 1.1 summarises a long exhaustive list of completed and ongoing projects funded by the European Commission (EC) since 2005.

While progress of technology is steadfast with several projects ready to go on the market, the biggest hurdle somehow at this point lies within the main regulatory framework and the legislation bodies. The stupendous advancement of technology foresees the danger of outstripping existing laws on mobility and transportation, with some of them dating back to the era of horse-drawn carriages. Therefore, if autonomous driving is to become an

unavoidable reality in the near future, policy makers should react quickly to anticipate the good functioning of this generation of cybercars.

1.2 This Thesis

Our aim is concentrated around building ego-centric topometric maps represented as a graph of keyframe nodes which can be efficiently used by autonomous agents. The keyframe nodes which combines a spherical image and a depth map (augmented visual sphere) synthesises information collected in a local area of space by an embedded acquisition system. The representation of the global environment consists of a collection of augmented visual spheres that provide the necessary coverage of an operational area. A "pose" graph that links these spheres together in six degrees of freedom, also defines the domain potentially exploitable for navigation tasks in real time. As part of this research, an approach to map-based representation has been proposed by considering the following issues:

- How to robustly apply visual odometry by making the most of both photometric and geometric information available from our augmented spherical database
- How to determine the quantity and optimal placement of these augmented spheres to cover an environment completely
- How to model sensor uncertainties and update the dense information of the augmented spheres
- How to compactly represent the information contained in the augmented sphere to ensure robustness, accuracy and stability along the trajectory by making use of saliency maps

This research work builds on the back of the successful CityVip project and extends the results presented in [Meilland *et al.* 2010], [Meilland *et al.* 2011a], [Meilland *et al.* 2011b] in several ways. With the aim of robustifying Visual Odometry, a cost function is introduced combining both geometric and photometric constraints applied in a direct image registration framework for pose estimation. Additionally, the problem of pose graph optimization is addressed, whereby, given a database of augmented visual spheres, an explored trajectory with redundant information is pruned out to a sparse skeletal pose graph.

Mapping under the assumption that the environment is static is somewhat baseless since the environment under which the robot is normally deployed is dynamic and evolves in unpredictable ways. Though some parts of the surrounding are static in the short run, others may be changing abruptly due to activities occurring around the robot– people walking around, cars on the road. This aspect is treated in our work by introducing the notion of dynamic entities which evolve along a trajectory. In addition to the initial graph of augmented spheres comprising of the photometric, geometric and the saliency map, two

more components are now tethered to the environment information content of our graph; the uncertainty map and the stability map.

1.2.1 Manuscript organisation

Chapter 2 lays down the state of the art of vision only slam systems. Limitations of other sensing technologies such as the GPS, wheel encoders, laser range scanners have broaden the horizon of researchers by exploiting the rich contents of images. The historical roadmap of visual odometry (VO) showed that initial efforts were concentrated towards building robust mobile robots for planetary explorations. Later on, visual odometry techniques became more diversified with applications to terrestial, aerial, sub-sea intelligent vehicles. Augmented reality is another promising area which requires VO. VO as other odometry techniques is not spared by the problem of drift. Two different approaches exist in literature; feature-based and dense-based. Both techniques are highlighted with their benefits and inconveniences. Further examples are detailed from literature to show how VO fits the SLAM framework. For localisation purposes, a robot normally requires a built map of its perceived environment. This can be done on the fly during the exploration phase or a map representation can be constructed separate from the exploration task during a first learning phase. In this context, the mapping framework is made up of several layers; metric topological, semantic. The pros and cons of each one of them is elaborated. Finally, the emerging field of lifelong mapping is worth a mention. This area aims at equipping intelligent vehicles with new tools to tackle the challenges of vast scale exploration with limited resources such as memory capacity, computational power or fully autonomous unsupervised navigation. Moreover, the robot should have the ability to create a stable map which can be used over and over again for long periods of time.

Chapter 3 introduces the reader to the world of 3D Vision. In order to illustrate the whole process of image formation onto the camera frame, a direct application involving camera calibration is discussed. To further extrapolate 2D images to 3D information as perceived by human vision, the concept of stereo vision is outlined, describing all the intermediary stages until the depth information extraction. These introductory concepts are vital to understand the extension to spherical vision, developed in the second fold. The motivation sparks from the multitude advantages it offers; a compact but enriched environment model with 360^0 omnidirectional field of view (FOV) representation as well as its invariance to rotation making it indifferent to sensor configuration. Moreover, covering the explored environment with the maximum possible range and orientation leads to better localisation. These aspects have encouraged researchers to come up with innovative ideas such as catadioptric cameras or better, multi baseline camera systems. While in the former case, a spherical image is readily obtained, for the latter case, obtaining spherical panoramas are not straightforward. This requires intermediary operations on images such as warping, blending and mosaicing. In this chapter, an overview of these techniques is provided with application to indoor and outdoor multi sensor systems developed as part of the

research activities of our team. For each system, its corresponding calibration technique is overlaid which outputs the extrinsic parameter matrices for the multi camera system. This step is vital for producing synthesised spherical panoramic images as it shall be seen.

Chapter 4 is based on spherical RGB-D odometry. It begins with an understanding of the optical flow model and extrapolated to 3D scene flow as a direct application to motion estimation. This important concept forms the backbone of the Lucas-Kanade's direct image based registration technique where a relative unknown parametrised motion between two image frames with small interframe displacement is computed. An overview of the photometric cost function is function is given as applied to our set of spherical augmented spheres consisting of spherical RGB images and their corresponding depth and saliency maps. Intensity based techniques show their limitations in poorly textured areas or regions subjected to high illumination variations. To compensate the weakness of such methods, a second cost function is introduced where the geometric information content of the depth map is explicitly taken into account with the implementation of a point to plane iterative closest point (ICP) technique inspired from literature. This leads to an improved formulation incorporating both geometric and photometric information in a hybrid minimization cost function. The global environment is represented by a pose graph consisting of nodes and edges, established from spherical VO, whereby each node is represented by a keyframe. Keyframe selection is a vital task in such a representation as careful selection of frames results in several advantages. Firstly, data redundancy is minimised, hence rendering the representation sparser and more compact. These are highly desirable properties when exploring vast scale environments with limited capacity. Secondly, using less keyframes helps in reducing the integration of tracking drift errors emerging from frame to keyframe odometry. Along this line, two different criteria are considered, notably, the median absolute deviation (MAD) and the other based on differential entropy which is an abstraction of the pose uncertainty. To validate the several aspects discussed in this chapter, a results section elaborates on two synthetic and two real datasets. The strengths and weaknesses of the algorithm are exploited before the conclusion is wrapped up. The experimental evaluation of this chapter was partly published in one international [Gokhool *et al.* 2014] and one national [Rives *et al.* 2014] conferences.

Chapter 5 tackles the various shortcomings exposed in the previous chapter, notably, the problem of drift or odometry failures which disrupts the global consistency of our pose/feature graph representation. A first approach seeks to improve the noisy depth map output from the sensor. In order to fuse depth map measurements, they should be represented in a common reference frame. Observations acquired along the trajectory are transferred to that particular frame by an inverse warping transformation. To detect inconsistencies arising due to occlusion phenomena or noise between the measured and the observed depth maps, a statistical test is implemented. The latter is principally based on drift detection on the mean time series of discrete signal. The limitations perceived during the evaluation of this first approach led to the consideration of a second methodology which

takes into account random and dynamic errors coming from the sensor. A much improved uncertainty model is formulated considering both the design of our spherical representation coupled with uncertainties coming from the warping operation. A probabilistic data association technique is devised in order to detect outliers coming from noise, occlusion, disocclusion and free space violation phenomena. Observation values in agreement with the measurement are fused so as to improve the geometric and photometric content of the augmented spheres. The aspect of dynamic points are further treated leading to the emergence of a new entity in our set of augmented spheres which is the stability map. This additional attribute is used to update the saliency map. An experimental section evaluates this second approach before the chapter is concluded. Part of the conceptual formulation and experimental evaluations of this chapter has been published in [Gokhool *et al.* 2015].

Chapter 6 summarises the evaluation of this work and perspectives are discussed to give an idea of how to take this mapping project to the next level by further exploiting the underlying VSLAM framework.

1.2.2 Publications

International:

- T. Gokhool, R. Martins, P. Rives and N. Despré. *A Compact Spherical RGBD Keyframe-based Representation*. In International Conference on Robotics and Automation, (ICRA), Seattle, US, May 2015.
- T. Gokhool, M. Meilland, P. Rives and E. Fernàndez-Moral. *A Dense map Building Approach from Spherical RGBD Images*. In International Conference on Computer Vision Theory and Applications, (VISAPP), Lisbon, Portugal, January 2014.

National:

- P. Rives, R. Drouilly and T. Gokhool. *Représentation orientée navigation d'environnements à grande échelle.* Reconnaissance de Formes et Intelligence Artificielle, RFIA 2014, France, June 2014.

CHAPTER 2

State of the Art

2.1 Introduction

Visual Odometry (VO) is defined as the process of estimating the relative motion of a mobile agent using vision sensors. This incremental technique computes the pose of a vehicle based on the movements induced by onboard cameras. Over the years, VO has been useful to compensate other similar techniques such as wheel odometry which is highly affected by dead reckoning in uneven terrains. On the other hand, global positioning system (GPS) has shown its limitation in aerial, underwater applications. In urban canyons type of environment, multiple reflections of GPS signals from skyscrapers' façades provide inaccurate measurements. As of late, breathtaking advancements in both vision sensors and computing hardware have made computer vision algorithms more mature down the years. Consequently, efficient, low cost odometry solutions provided by vision systems have seen widespread applications in the field of mobile robotics.

This chapter provides a broad coverage of the approaches and methodologies involved in equipping mobile robots with vision technology, starting from the theory of VO, through intelligent modelling of the environment before ending on the techniques for lifelong autonomous deployments.

2.2 VO & SLAM

Historical roadmap The first footprint of egomotion estimation applied to mobile robots appeared in the early 80s with the work of [Moravec 1980]. The efforts in this era were particularly motivated around the conception of planetary rovers and provide them with the ability to measure their six degree of freedom motion in unstructured rugged terrains where wheel odometry is highly unreliable, notably due to wheel skidding. In this work, a single camera was made to slide on a stop and go motion style, digitizing and analyzing images at every location. This approach termed as *slider stereo* computed the rigid body motion by aligning 3-D points perceived between two consecutive frames. Between each stop, the camera slid horizontally on the rail, taking 9 snapshots at known equidistant intervals. A novel corner-based feature extraction technique was used to extract image primitives in a reference frame which was matched along the epipolar line of the 8 others using normalised

cross correlation as similarity criteria. Outliers were removed using a coarse to fine strategy accounting for large scale changes and the system was solved in a weighted least means square fashion to obtain the relative pose.

Although a single camera was used, this work belongs to the category of stereo VO simply because 3D positions are directly measured by triangulation, as do trinocular methods as well. One typical drawback is that motion can only be recovered up to a scale factor. The absolute scale can then be determined using motion constraints or from measurements obtained by using additional sensors such as IMUs and range sensors. Even then, monocular methods attracts particular interest since VO degenerates into monocular for the case where the distance to the scene is much larger than the stereo baseline as pointed out in [Scaramuzza & Fraundorfer 2011].

Since then, the framework of [Moravec 1980] has been adopted as a guideline for further improvements. The binocular system of [Matthies 1980] incorporated feature uncertainties in the pose estimation phase, which even saw better results in terms of relative trajectory error obtained. Later on, along the same streamline of developping planetary mobile robots equipped with VO technology, [Lacroix *et al.* 1999] focussed on good pixels to track. Reliable pixels are those whose corresponding 3D point are accurately known based on a stereo vision error model. Additionally, Pixel selection is further enhanced by studying the behaviour of key pixels around their neighbourhood. Interestingly, the correlation score curve extracted around that key pixel was found to have a strong relationship with disparity uncertainty – the sharper the peak observed, the more precise is the disparity. The neighbourhood pixel information was further used to discard points which may drift over subsequent frames due to false correspondences occurring in low textured areas. Pixel tracking between a stereo frame T_0 and T_1 is done by finding correspondences in the search zone of image frame T_1 predicted by an estimated transformation uncertainty. This prediction is important to restrict the search zone thereby increasing the chance of finding an inlier, or better, rejecting an outlier match. Ultimately, egomotion estimation is computed based on a 3D–3D constrained weighted least square approach proposed in [Haralick *et al.* 1989]. However, using wheel odometry feedback for an initial pose estimation is not always reliable. How the problem of slippage was tackled was not discussed by the author.

The conceptual analysis made in [Lacroix *et al.* 1999] was later implemented on the Mars rover platform later published in the work of [Cheng *et al.* 2006]. The only difference is that motion computation is encapsulated in a *Random Sample Consensus*, RANSAC mechanism [Fischler & Bolles 1981]. RANSAC is an established model fitting paradigm to perform parameter estimation taking into account noisy data. Given two initial sets of data points randomly sampled, one pertaining to the source and the other, to the destination set, the objective is to obtain the best model parameters. For the case of VO, the hypothesized model is the relative motion (R, t) between two camera frames and data points extracted from these two sets are then candidates for feature correspondences.

The term *Visual Odometry* was epitomised in the landmark paper of [Nistér *et al.* 2004] though research in this field has been ongoing for the last two decades or so, leading to his

real time VO algorithm. Feature detection was based on a Harris corner implementation of [Harris & Stephens 1988] using an optimal number of operations leaning on MMX coding instructions. Robust motion estimation is computed using in an iterative refinement fashion using a preemptive RANSAC [Nistér 2003], a variant of the original algorithm. To validate the method, a mobile robotic platform was integrated with the VO component. The vehicle was also equipped with a *Differential* GPS, (DGPS) as well as a high precision *Inertial Navigation System* (INS). The highly accurate INS/DGPS system was used only as a comparative ground truth. The mining-like truck vehicle was made to navigate in rough agricultural and forestial environments. Results obtained were impressive, with a positional error as small as 1.07% with respect to the DGPS and an almost negligible orientation error as of around 0.6^0 with respect to the INS on a track of around 600m. This work is considered as a genuine VO breakthrough for autonomous ground vehicles.

2.2.1 Feature based vs Dense based

Odometry techniques in general require accurate relative motion estimation to reduce trajectory drift. VO, which relies heavily on image contents requires at first hand good quality feature matching which makes the problem difficult [Fitzgibon 2003]. An important step prior to registration requires that data coming from two viewpoints should be put in correspondence. Two main approaches are identified; one which goes through an initial feature identification phase between the two data samples while the other uses dense correspondence technique.

Feature based This approach is aimed at extracting salient points in an image which are most likely to find a good match in other images. For example, a blob is an image feature whose characteristics such as intensity, colour and texture differ distinctively from its immediate surrounding. On the other hand, a corner is a point occurring at the intersection of two or more edges. Good features to track must generally possess specific properties such as invariance to illumination, rotation, scale or perspective distortion. They must be robust to noise, compression artifacts or blur. They should be easily redetected in other images, thus repeatable. They must provide accurate localisation both in position and scale and finally, they must be notably distinct so that correspondences can be found in other images, especially, those covering the same ground at two different vantage points. In quest of finding features with the above mentioned properties, a great deal of effort has been input by the research community. Popular feature detectors are Harris [Harris & Stephens 1988], Shi-Tomasi [Tomasi & Shi 1994], FAST [Rosten & Drummond 2006], SIFT [Lowe 2003], SURF [Bay *et al.* 2006], BRISK [Leutenegger *et al.* 2011], ORB [Rublee *et al.* 2011] to name a few. An indepth analysis of each one of them is provided in [Fraundorfer & Scaramuzza 2012].

Similarity measures such as the *sum of squared differences* (SSD), *sum of absolute differences* (SAD) or the *zero centred normalised cross correlation* (ZNCC) are common

criteria used to evaluate matches across images. Feature matching techniques can be further subdivided in two further approaches. In the first, extracted features from the first image is matched across subsequent frames using local search techniques based on correlation as discussed above. This works well for close sequential images in a video for example. In the second approach, two sets of features are extracted in two images located at different viewpoints and are then matched for correspondences. This is more suited for wide baseline camera frames where images are taken from two completely different viewpoints resulting in a direct application where large scale environments are involved. Such techniques help to overcome motion drift related issues.

Feature based techniques are frequently used for egomotion estimation since they provide a compact information content as compared to using all the image pixels. However, they rely on an intermediary estimation process based on detection thresholds. This process is often ill-conditioned, noisy and not robust thereby relying on higher level robust estimation techniques as pointed out in [Meilland & Comport 2013b].

Dense based Also known as direct method, this technique does not require preprocessing of 2D points for matching. Instead, the entire content of a source image $\mathcal{I}^*$ and a destination image $\mathcal{I}$ are used to compute the camera motion in a direct intensity minimisation approach embedding a parametrised motion. Indeed, this minimisation function is non linear, which requires solvers provided by *Gradient Descent*, *Gaussi- Newton* or *Levenberg-Marquardt* class of optimisation tools. This technique works well for very small interframe incremental motions but due to the fact that the whole image information content is used, the minimisation approach takes advantage of the massive data redundancy to output a more robust and precise pose estimate. Initially proposed by [Lucas & Kanade 1981], several variants later appeared in literature improving the original algorithm in terms of computational cost and robustness [Baker & Matthews 2001] , while [Malis 2004] later provided an even better conceptualisation of the method based on a second order approximation of of the linearised objective function. Though computation is increased, the methodology provides better convergence properties as well as more robustness at the solution.

Once the pose has been recovered, an updated dense depth map can be obtained by using the same intensity cost function but this time the estimating depth becomes the parameter to be minimised in a *Maximum Likelihood Estimation* MLE fashion. However, this technique produces erroneous surface estimates whenever the Lambertian assumption is violated with dynamic lighting conditions in the scene as well as partial observability conditions across multiple views. To obtained a denoised depth map, the problem is converted in a *maximum aposteriori* (MAP) function which includes a noise model. The better consistent depth map is then obtained using regularization techniques as elaborated in [Newcombe 2012]. Initially limited to 3D object reconstruction due to expensive regularization, the approach is gaining ground in dense scene reconstruction applications due to the recent advancements in computational power [Pizzoli *et al.* 2014].

2.2.2 Visual SLAM

GPS localisation can be accurate as 1 cm, for e.g: the Real Time Kinematic (RTK) GPS, but under the conditions of sufficient satellites' visibility from the receiver. However, in densely populated areas, such as urban canyons, or indoor settings, accuracy drops considerably. In this context, the group of Maxime Lhuillier of LASMEA proposed an alternative based on visual monocular SLAM where vision sensors are explored. The related work of [Royer *et al.* 2007], detailed the concepts of offline map building as well as its attributed advantages. However, rapid accumulation of camera frames along a given trajectory leads to the piling up of redundant information which weighs heavily on the allocated memory. To cater for that, keyframes are selected according to a defined criteria in order to obtain an optimised pose graph - other methods such as the one mentioned in [Nistér 2001], is based on batch processing of image frames where the redundant ones are eliminated or the Geometric Robust Information Criterion (GRIC) proposed by [Torr *et al.*].

Furthemore, an improved pose performance can be obtained by the fusion of the covariance observed from the dynamic model of the vehicle with the optimisation process. Image matching, keyframe selection, pose uncertainty estimation and propagation are further treated in their works [Mouragnon *et al.* 2006b][Mouragnon *et al.* 2006a]. Uncertainty propagation has been further detailed at length in [Lhuillier & Perriollat 2006]. Their formulation rests on the backbone of Structure from Motion (SfM) algorithm resulting in the implementation of variants of bundle adjustment (BA) optimisation. BA as defined in [Triggs *et al.* 2000] is a minimisation problem based on the sum of the squared reprojection errors between the camera poses and 3D point clouds.

Constantly changing features in unstructured dynamic environments lead to the problem of unstable mapping with little reusability, for e.g, urban environments, some landmarks are fixed such as buildings, roads and sideways while other features are dynamic in long term - vegetation, impacted seasons, or short term - vehicles in locomotions, billboards, pedestrians which usually make up the complete real life picture. [Royer 2006] suggested that maps should be updated each time the trajectory is revisited so that new features may be added up and old ones with no sign of observability are discarded.

Real-Time Visual SLAM can be diversified into two distinct successful approaches; filtering and keyframe methods. The former fits into a SLAM framework where the states related motion of a mobile robot are estimated in real-time as the robot continously observes an explored environment. On the other hand, the latter emerges from the SfM research area with BA optimisation being the main focus. However, being it BA in SfM or EKF based probabilistic SLAM, both of them rest on the backbone of Gaussian probability distribution theory because of its ease to represent the notion of uncertainty measurements and estimation. In the filtering approach, only the current pose is retained at the expense of the previously recorded history while pertinent features with a high predictive probability are maintained. Alternately, BA optimisation approach involves solving the graph from scratch each time with the incoming frames but at the same time, discarding redundant keyframes which contribute little to estimates. An in-depth synthesis of the methods

described in the paragraph above, has been treated at large by [Strasdat 2012]. He further argues that though the number of features in the graph are very high, this technique of BA is more accurate and stable over systems using joint estimation with uncertainty over sparser maps using a filtering framework. Perhaps the best references described for the two paths above are the works of Eade and Drummond [Eade & Drummond 2007] for locally filtered maps and the works of Klein and Murray [Klein & Murray 2007] based on Keyframe BA. The author further compares the factor of computational cost which grows linearly for the number of features in the case of BA and cubic in the case of filtering. A close up of the comparative study described above reveals that BA is far more superior than filtering in terms of robustness, accuracy and efficiency when the problem of large scale mapping is treated.

Parallel Tracking and Mapping (PTAM), proposed by [Klein & Murray 2007] was among the first work to broaden the horizon of VSLAM by demarcating from state of the art filtering techniques of [Davison *et al.* 2007, Eade & Drummond 2007]. The first novelty of the approach was the decoupling of tracking and mapping in separate threads. Tracking is performed from a coarse to fine level by first projecting an initial map model onto the current camera frame where feature correspondence is done using a patch search technique with the helpf of motion prior, using a constant velocity model. At the coarsest level, a small number of features (around 50) is used to find an initial estimate by minimising over the translational component only. The output of this coarse motion estimate is used to find more potentially visible image patches (around 1000) and further used to refine the pose estimate over a second robust intensity reprojection error function, with this time all 6 DOF is estimated. The second contribution is the use of keyframes . In order to avoid redundancy over accumulated images at frame rate, a set of heuristics is defined to preserve only meaningful data. Consequently, the map is made up of sparse keypoints stored in a keyframe-based representation referenced in a world coordinate system. BA runs at two levels; notably local and global implemented on a separate back-end thread as the one of tracking. The system was designed for augmented reality applications with a user in the loop to initialise mapping. Comparison with EKF SLAM of [Davison *et al.* 2007] showed better tracking performance. However, as new keyframes are added up, the system slows down (beyond a hundred keyframes) and eventually saturates, causing BA failures. Tracking also failed when an erroneous pose estimation goes undetected by the system thereby incorporating wrong information in the map. Nevertheless, the mapping system designed was one of the major leaps in VSLAM whereby encouraging results obtained opened up new perspectives. The shortcomings of this work were addressed in [Klein & Murray 2008].

Owing to the spectacular technological advancements made by the gaming industry recently, commodity RGB-D cameras attracted the interest of the robotics and vision community. As a result, RGB-D cameras such as the Microsoft Kinect or the Asus Xtion Pro Live are under exploration for indoor mapping assignments. The work of [Henry *et al.* 2012] presented a complete VSLAM system for a building-size environment. A front-end SLAM framework incorporating both texture and depth was used for frame to frame alignment.

Trajectory drift resulting from noise and quantisation errors is corrected by applying global optimisation at loop closure detection. One of the highlights of this work is the use of a surfel map representation. A sequence comprising of 95 images outputting more than 23 million points are reduced to around 730K surfels, accounting for a net reduction by a size factor or 32. Surfels (from surface pixels) encodes location, orientation, patch size and color of a particular surface. One important result is that their RGB-D ICP framework outperforms either RGB or ICP only alignment. Feature based RGB alignment results in greater mean distance errors wth a "night dataset" where poor estimation is obtained due to bad feature extraction in dark or textureless areas. ICP only copes well with such situations but ICP-only performance is below that of their hybrid framework. The take home message is that, by making an optimal use of both color and depth map outputs of the sensor, a consistent reduced drift environment map is achievable.

2.3 Map representation

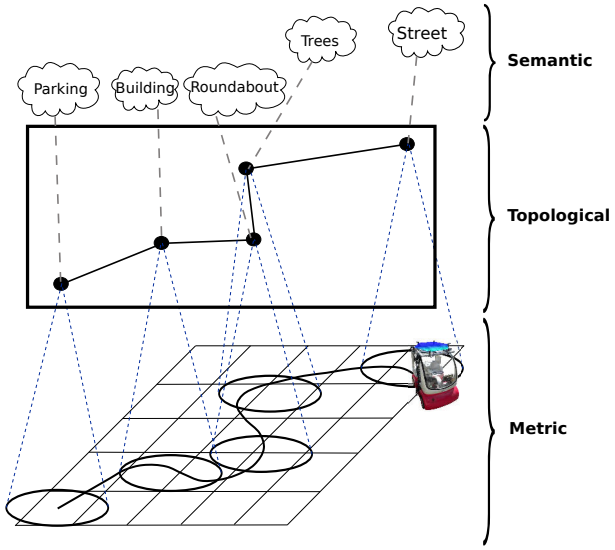

Figure 2.1: *Typical layers of a mapping system, courtesy of [Chapoulie et al. 2013]*

Map building is a registration process of the observed elements in a region. Several variants include evidence grids, point clouds, meshes or topological graphs which are common representatives of 3D maps. In this section, some typical map representations will be discussed, highlighting their pros and cons.

2.3.1 Occupancy grids

Categorised as a metric map, the occupancy grid [Elfes 1989] emerges from the geometric structure of the environment. For the case of a robot moving on a flat surface, the representation is in 2D whereby the region under exploration is partioned into evenly spaced

cells, also known as grids. Thus, the space can be viewed as a matrix of cells, whereby each cell stores a probabilistic estimate of its state bounded by $(0, 1)$ such that 0 indicates unoccupied empty space and 1 means definitely occupied space, while unexplored space is assigned a prior probability of 0.5. The representation is not obtained in a direct straight forward manner, since the sensory measurement provided by the robot needs to be churned out into a spatial world model. A stochastic sensor model is normally defined at the base of this data interpretation process where the measurement model is often approximated using a *probability density function* (pdf) relating the reading to the parameter space. To determine the state probability of grid cells, the pdf is then used in a Bayesian estimation procedure whereby the deterministic world model is extracted from a set of decision rules based on optimal estimators such as the *maximum aposteriori* (MAP) estimate.

The first major drawback in this representation is undeniably scalability over large extended environments. Finer details of the environment are captured at the expense of the grid resolution (also grid granularity) which leads to a greater computational and storage capacity exertion. Moreover, the larger the grid size, the more expensive the search becomes for data association. Adding up to that, tasks such as trajectory planning become computationally expensive for finer grid resolution. Therefore, it is needed to strike a right balance between granularity and computational complexity.

An Octomap [Wurm *et al.* 2010], is a direct extension of an occupancy grid to a 3D space model . It consists of an octree data structure which is basically a tree with nodes, with each parent node splitting into eight equal-sized voxels. The leaves of the tree contain an occupancy space probability at minimum resolution size allowing a compact environment representation whilst accounting for sensor uncertainty measurement. Occupancy of a voxel is further clamped between an upper bound and a lower bound such that a "hit" is observed when a tracing ray ends up in the voxel while a "miss" means that the ray traversed through the voxel.

In [Steinbrücker *et al.* 2013], a surface is stored as a volumetric *truncated signed distance function* (TSDF) in a 3D octree data structure. TSDF [Newcombe *et al.* 2011] [Levoy *et al.* 2000] is a discretisation technique used in dense surface reconstruction in order to finds an accurate surface prediction from noisy sensor measurements. It encodes a zero crossing ($+$ve to $-$ve) such that when a ray is cast from the camera centre to the object in space, each pixel of the raycast is marched starting from the minimum depth. Marching continues to be assigned positive as long as the pixel depth is infront of the object and stops when the backface is reached indicating a signal transition. [Steinbrücker *et al.* 2013] implemented a multilayer octree such that depth values are encoded from a coarser to a finer level depending on the depth uncertainty measurement which varies quadratically to the depth. This allows an efficient sparse representation of the map. Expensive update is performed only on the observed part of the map. Though a highly efficient GPU implementation is made, achieving better reconstruction quality outperforming state of the art on several datasets, this technique remains limited to building size type of environment.

2.3.2 Feature maps

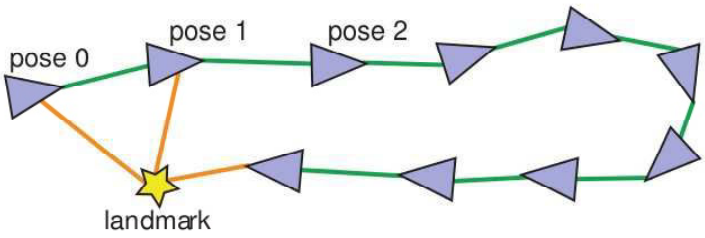

Figure 2.2: *Pose/feature graph, courtesy of [Olson 2008]*

Geometric primitives such as points, lines are integral components of feature maps. The map is built out of robot poses and perceived landmarks of the environment. A pose is by a robot position at a particular point in time and represents a node in the graph. In a similar way, landmarks are also represented as nodes. Following sensor measurement at a particular location, landmarks are linked to poses through edges depending upon the observability conditions across nodes (*cf.* figure 2.2). The terminology of "pose/feature graph" comes from the constituency of the mapping system. A graph containing both poses and features is related to a pose and feature map. If only robot poses are considered with edges linking them, only a "pose" graph emerges out [Olson 2008].

Landmark locations in an apriori feature map are assumed to be perfectly known and each feature is defined by its parameter set constituting of its 3D location, plus other attributes describing particular characteristics such as curvature, radius for example. Principally, feature location is useful for localisation while supplementary descriptive information helps for data association. Localisation is effected on first hand through a data association process where extracted features from sensor data are matched with those of the map. Discrepancies between the predicted and measured feature locations are then used to compute the vehicle pose. Similar techniques as the ones described in VO above can be used for localisation. Otherwise, filtering techniques such as the *Extended Kalman Filter* EKF can also be applied in a SLAM framework [Durrant-Whyte & Bailey 2006]. Recursive EKF pose estimation offers the advantages of efficient data fusion from multiple sensor measurements as well as the ability to incorporate sensor uncertainty models.

Feature maps offer a sparse representation collectively represented by landmarks and robot poses, unlike occupancy grids which ask for greedy dense description. However, due to the fact that free space is not represented, feature maps do not provide solutions for trajectory planning or obstacle avoidance tasks. These operations must be performed as a separate option. Moreover, data association is arguably, the main weakness of feature map localisation. Reliable pose estimates result from successful correspondences between feature observations and their associated map feature. Misassociations results in erroneous pose estimates leading to robot localisation and map update failures. Another loophole for such a representation is that feature maps owe their suitability only to environments where geometric feature models are conveniently extracted, which might not be necessarily the

case in unstructured environments where geometric primitives such as lines or points are hardly parametrisable. Hence, there is the need to devise parametric models that adequately describe objects for consistent extraction and classification.

Nevertheless, they offer a very compact representation of the environment which explains their exploitation. [Fairfield 2009], treated this area of work based on sonar sensory measurements for underwater exploration. 3D evidence grids are overlaid on a novel approach termed as the Defferred Reference Count Octree (DECO) that exploits the spatial sparsity of many environments. Evidence grids are based on a probabilistic approach while mapping matching consisting of two distinct sets of range data is performed using an Iterative Closest Point algorithm (ICP). Finally, a map-alignment algorithm has been adopted from the classic Lucas Kanade method with a flavour of Baker and Mathews's [S.Baker & Matthews 2001] Inverse Compositional (IC) for computational efficiency.

2.3.3 Topological maps

Topological maps are rather contrasting propositions with respect to the mapping techniques discussed in previous sections. They do not rely on metric information as in the case of occupancy grids and feature maps. Instead, a graph structure consists of nodes defining distinct places of the environment joined together by edges representing the accessibility between places generating the graph. The workability of the concept rests upon assumptions that distinctive places are locally distinguishable from their surrounding area and the procedural information is sufficient for the robot to travel to a specified location with a recognising distance.

Feature detectors (*e.g.* SIFT, SURF, FAST) are used to extract visual words from images and stored on the fly in a codebook made up of visual vocabularies and inverted index files [Chapoulie *et al.* 2011]. Localisation is purely appearance based where visual words of the image of the current viewpoint are matched in the dictionary of words in order to retrieve the most likely image of the database. For place recognition to function correctly, a node description must be unique along the connecting path regions from its adjacent nodes. In this context, [Chapoulie *et al.* 2012][Chapoulie *et al.* 2013], provide a topological segmentation based on change detection in the structural properties (textures, appearance frequency, orientation of straight lines, curvatures, repeated patterns) of the scene during navigation. On the whole, topological representation not only brings about a good abstraction level of the environment bu common tasks such as homing, navigation, exploration and path planning become more efficient.

Their fundamental weakness resides in the lack of metric information rendering some of the above-mentioned tasks precision deficient as only rough location estimates are available. Travelling between nodes using purely qualitative trajectory information might work for static structure environments but rather inappropriate for more complex dynamic scenarios. Perceptual aliasing (also false positive) further adds to the weakness of this representation. This happens when two distinct portions of the environment appear similar,

which is much frequent in highly structured environments (*e.g.* offices with apparently similar cubicles). Fortunately, with the higher level of information abstraction for vision based systems, this problem can be mitigated. On the other hand, false negatives occur due to places undergoing modifications, which can be natural or man made. These include viewpoint variations, occlusions, structural changes, dynamic objects, lighting conditions, to name a few. Ultimately, both geometric and visual recognition methods are sensitive to these forms of failures [Bailey 2002].

2.3.4 Topometric maps

Summing up, the characteristics of metric and topological maps are complementary. To harness the benefits of each one of them, they should be included within the same framework. Metric maps give the notion of uncertainty in the representation which allows data association to be made within a confidence interval. On the other hand, topological maps give a sparser representation by breaking the world into locally connected regions. Topological maps generally sit upon the metric layer at a higher level of abstraction (*cf.* figure 2.1). Localisation is performed sequentially, first on a topological appearance based level before obtaining the exact pose of the robot through the metric layer. The main focus of topometric maps is to create accurate metric submaps adhered to distinct places of the environment whilst covering the sensor range of the robot. Concrete examples of such implementations may be found in [Dayoub *et al.* 2013, Badino *et al.* 2011].

2.3.5 Semantic Maps

These days, the field of mapping is undergoing a considerable shift of paradigm. The objective is not only to build representations simulating appearance and 3D space of the environment but also to enrich the information content with a touch of "human based" scene understanding. The task is to model the semantic content of the environment with objects it contains to take into account phenomena that are otherwise ignored by metric and topological representations. Semantic mapping offers a rather natural means of information sharing in a similar way to the level of human understanding rendering human robot interaction more efficient and simpler, attributes which are essential for robots deployments' in our day to day life. Furthermore, understanding the nature of objects give the possibility to model interactions with them. This allows us to consider the dynamic aspect of the scene by associating a certain type of behaviour to these objects. For example, defining object classes which possess good landmark attributes (*e.g.* road signs) can help enormously to adapt navigation tasks. Identifying specific objects in certain regions may provide a better region segmentation (*e.g.* dishwashers, dining table, cuttleries are most likely to be found in a kitchen). Semantic maps augments the mapping structure (*cf.* figure 2.1) by an additional layer of abstraction pertaining to scene understanding and spatial reasoning. This newly emerging area offers promissing perspectives in terms of large scale navigation and mapping as discussed in [Drouilly *et al.* 2013].

One of the most complete model in recent times is provided by [Pronobis & Jensfelt 2012]. The mapping system consists of four different hierarchical layers; metrical, topological, categorical and conceptual. The first two layers have already been discussed in previous sections. The categorical layer contains models describing objects and landmarks as well as spatial properties such as geometrical models of room shape or a visual model of appearance. Laid upon the the conceptual layer is the categorical layer which is the core of this work and replicates human-based reasoning. For example, an oven, a dishwasher, an air extractor are objects more likely to be found in a kitchen rather than in an office. The learning phase is based on a heavy machinery of machine learning tools such as *support vector machine* (SVM) for data classification and *conditional random field* (CRF) for pattern classification. This work is considered as one of the first footprints of high level semantic mapping.

An altogether different proposition is made in [Salas-Moreno *et al.* 2013], where high quality 3D models of repeatedly occurring objects are first learned manually offline and then stored in a database. For online operation, 3D object recognition is achieved on a basis of correspondence voting between *Point-Pair Features* (PPFs). This involves finding metric consistencies between 4D descriptors consisting of relative position and normal pairs of oriented points on the object surface. Their object recognition module works in real time by efficiently exploiting the GPU and works well in cases where objects are partly occluded by using view prediction. However, this technique works well for limited indoor environments with repeated identical object elements. In this work, object segmentation is purely metric and far from the higher level of object abstraction and knowledge based reasoning of [Pronobis & Jensfelt 2012].

2.4 Towards Lifelong mapping

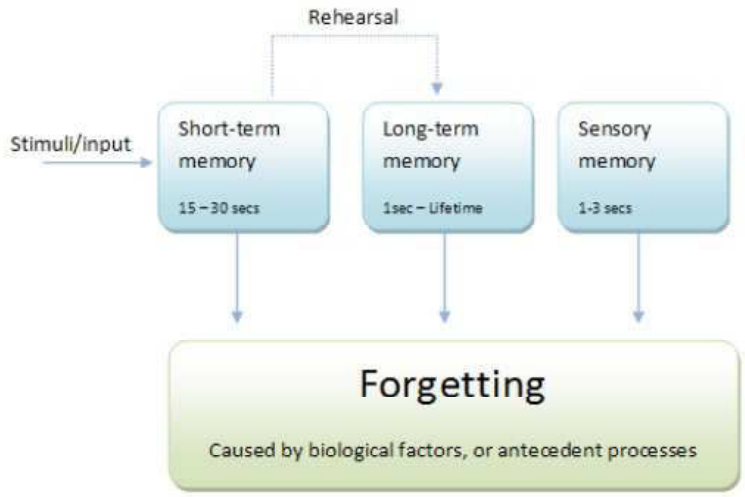

Figure 2.3: *Atkinson and Shiffrin human memory model*

2.4.1 Memory Models

The scientific community has always been looking for inspiration out of nature. As a matter of fact, the origins of the probabilistic localisation models have been drawn from the fundamental concept of "place cells " in the hippocampus. The [Atkinson & Shiffrin 1968] memory model is made up of four main components as shown in figure 2.3; the short term memory (STM) which retains information temporally but long enough to recall it; the long term memory (LTM) which retains information for extended periods of time or for lifetime; the sensory model (SM) integrated afterwards to accomodate for the functioning of the sensing organs as an input port to the signals of the external world which need to be processed. Finally, the forgetting module affects all the other components since it is attributed to the fact that memories can be forgotten through trace decay. Stimuli inputs enter the model through STM. If the inputs are continously rehearsed, they are transferred from STM to LTM. Even though information retained in the LTM is continously recalled in a lifetime, it is not guaranteed to stay permanently over there. Meaning that if it is not adequately rehearsed, it can be flushed out of memory. This memory model has lately been adapted in the work of [Dayoub *et al.* 2011] for robot mapping.

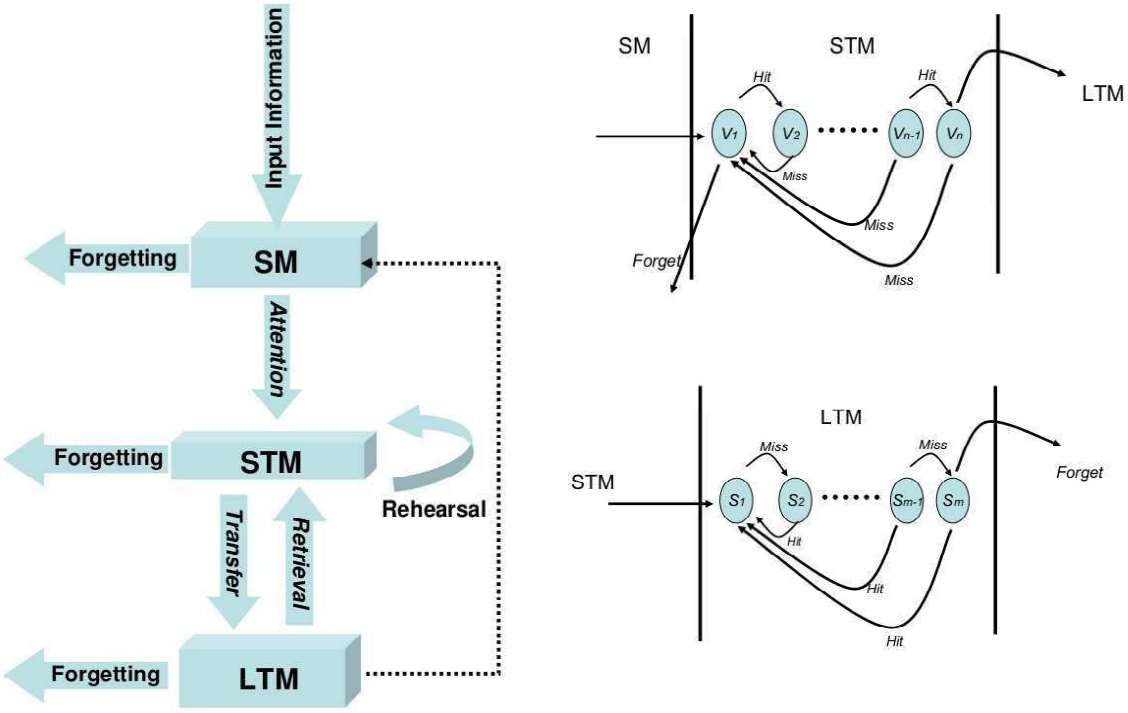

Figure 2.4: *Long term and short term memory model, courtesy of [Dayoub et al. 2011]*

In [Dayoub *et al.* 2011], the model was applied to overcome the limitations of metric and topological maps for environments which evolve over time. The environment was represented as an adjacency graph of nodes on a topological level, where each node was connected to edges on a metric level. The nodes represented distinctive places of the en-

vironment while the edges symbolises transitions between them. The approach consisted of two stages; in the first, the robot is unleashed into its operational space to learn an environment model and a map is built out of perceptual images acquired. In the second stage, the robot makes use of the map for localisation, *i.e.* to identify the most likely node to the current view as well as its relative metric transformation from that node. Memory activity starts as from the localisation process and from there on, the mechanism is triggered to update the map with new measurements as well as to flush out redundant information. After the raw sensory data has been processed into salient environment features, they are transferred into the STM and move across the STM channel through a recall, rehearsal and transfer steps. If a particular feature in the current view is found to be corresponding to that of the reference view, a "hit" is registered and the feature moves a step closer to the LTM. Only when that feature has been persistently registered during every data association phase that it is allowed to progress over to the LTM. If no association has been found, a "miss" is attributed to that feature and demoted across the STM channel until it is eventually forgotten. In this way, spurious features are easily rejected while persistent features are transferred to the LTM. The same process is repeated for the recall state of the LTM in a finite state machine representation as shown in figure 2.4. Data association is evaluated using a nearest neighbour matching scheme and updated with an unscented Kalman filter. To evaluate their methodology, several experiments were devised with simulated and real environments. The differences were quite significant since the number of matched points during localisation were around 77% for the case of the static map while this figure rose to 95% with the use of adaptive memory model. Better results were also obtained when operating in a dynamic environment and it was shown that their approach superseded that of using only a static map.

The work of [Dayoub *et al.* 2011] was recently extended in two folds. Firstly, in the work of [Morris *et al.* 2014b], the LTM and STM memory model are represented as two 3D Octomap representation using RGB-D point clouds. The LTM is first initialised using a prior mapping stage while the STM starts as empty. A set of policies are defined based on log-odds values in order to rehearse the content of the LTM and the STM. Multiple map hypotheses are obtained by projecting the LTM'S content into multiple layers. These distinct layers represent new updated information of the world, decaying information due to gradual changes and forgotten information resulting from changed environments such as blocked passages and doors opening or closing. Each map hypothesis are then evaluated through a navigation phase. A particular map hypothesis is evaluated by the localisation accuracy obtained by using a confidence measure on the pose estimates. The robot switches between different map hypotheses in order to plan the best trajectory towards a certain goal. The envelope of this work was further pushed in [Morris *et al.* 2014a]where these multiple map hypotheses are handled in a more intelligent way by considering the localisation influence on locally accurate odometry. The Kullback-Leiber divergence between two pose sets, one with the original integrated odometry and one perturbed by localisation is computed and the map hypothesis yielding the least diverging score is eventually selected for path planning.

Earlier, [Biber & Duckett 2009] introduces the concept of timescales in the context of long term SLAM. After a first exploration of a desired trajectory, an initial set of local maps is built using laser scans and odometry as input. Several runs are made to update a short term local map ,fissioned into multiple submaps associated to a timescale spectrum while the long term memory maps are pruned at a lesser frequency offline. However, the notion of the number of submaps which should be attributed to each timescale is not clear and this type of system results in a very large database which requires efficient data management.

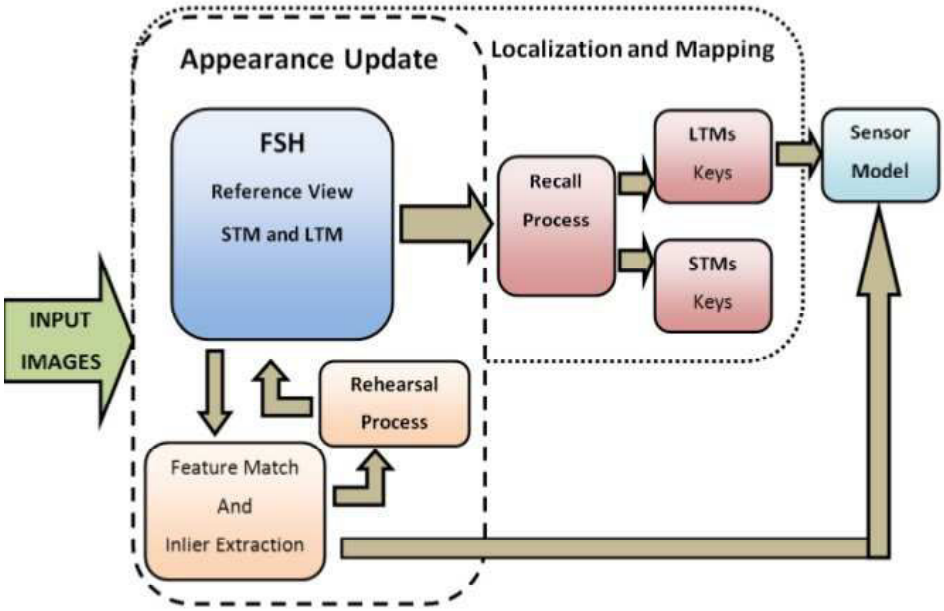

Figure 2.5: *Feature Stability Histogram memory model, courtesy of [Bacca 2012]*

The memory model proposed in [Atkinson & Shiffrin 1968] has been recently challenged by [Baddeley 2003][Llinas 2002] due to its linear representation of the memory process. They argued that this model does not take into account the ability of many people to recall information without it being rehearsed. Logically, it seems that stimuli inputs can bypass STM and land directly in LTM. Furthermore, the proposed memory model does not consider multiple memory layers. Consequently, in terms of robotics mapping, it would be useful to take into account memory levels rather represented by the strength of feature information. This issue is addressed in [Bacca 2012].

The reference view composed of the STM and the LTM has two main properties. Firstly, an input feature is allowed to bypass the STM and integrate the LTM based on its strength derived from its uncertainty value. Secondly, using a Feature Stability Histogram (FSH) in the reference view, feature classification into STM/LTM is non-linear since the rehearsal process takes into account the number of times a feature has been observed weighted by its corresponding strength value. The recall process classifies features into LTM and STM according to an upper threshold set in the FSH. Below that threshold, the feature is inserted in the STM. According to the author, such classification allows to

deal efficiently with temporal occlusions, dynamic environments and illumination changes. Eventually, only features belonging to the LTM are used for mapping and localisation.

Their approach was experimentally verified using localisation accuracy as a fundamental criteria and compared with two other well known techniques; that of [Dayoub *et al.* 2011] and the based on a Bag of Words (BoW) method of [Aly *et al.* 2011]. Both methods were outperformed by FSH, with better localisation accuracy and a lower uncertainty bound. However, the technique of [Dayoub *et al.* 2011] picks up later in the long run and rectifies the initial discrepancy with an increasing number of experimentations. But due to the rigidity of the update and rehearsal strategy based on a Finite State Machine (FSM) mechanism, the system suffers from delayed appearance update as it heavily depends on the number of states a particular feature has to undergo in order to make a successful STM to LTM transition. On the other hand, the BoW method followed a similar trend to FSH but with greater uncertainty bound and is more sensitive to natural illumination changes over seasonal changes. Two loopholes of the FSH are identified; the first lies in its deficiency to fuse visual features with their corresponding metric information and secondly, the disadvantage of the feature covariance suffering from overconfident uncertainty. The latter as aspect, needs consideration as backbone of FSM relies on feature strength. Overall, better quantitative and qualitative results are obtained compared to other state of the art approaches.

2.4.2 Graph pruning/ optimisation

Using a pose graph as a key environment representation for mobile robots constitutes of modelling robot poses coming from the spatial constraints between poses resulting from observation or odometry. Extraction of these constraints directly from sensor data forms the main component of front end SLAM. However, as observations are made, nodes and edges making up the graph builds up, memory as well as computational complexity of the mapping system becomes a central issue. Along that streamline, [Kretzschmar *et al.* 2010] proposed a novel approach in order to efficiently prune out graph nodes. Whenever an observation is made, its expected information gain is evaluated before deciding to insert a new node in the graph. The approach is meant to operate in the context of lifelong mapping of static environments. For tasks such as trajectory trajectory planning, many robotics systems require the need of the pose graph structure to be converted into other data structures such as feature maps or occupancy grid maps. In this work, the occupancy grid mao was the preferred choice of application of the novel information based node reduction algorithm. The author elegantly unveiled the application of information theory computed over the full map posterior in order to take into account both pose and map uncertainty. The added value of a particular observation is evaluated based on the entropy of the information gain. The idea is to ignore observations and their corresponding graph nodes whose expected information drops below a certain threshold. This work also include a mechanism for graph update without increasing its complexity when a node is removed. The experimental evaluation successfully demonstrated their approach. The robot was made to re-visit places forth and

back in a lab environment. During this exercise, other state of the art graph-based optimisation approach succumbed to run time explosion while the author's method avoided the run time explosion by keeping a constant number of nodes in the graph. The second result obtained dealt with the quality of the resulting grid map. When the robot re-visited the same places several times, due to scan-matching error accumulation, the structure of the observed environment is degraded by the misalignment of structures leading to artificial and blurred thick walls. Eventually, this problem is also avoided with their technique of pose graph sparsification. Finally, their claim of building a pose graph which only grows when genuine information value is acquired and not one which scales with the length of the trajectory was successfully verified.

In his thesis, [Olson 2008] discussed a method of hybrid map optimisation based on a Chi squared error function coupled with stochastic gradient descent method. The author stresses on the robustness of the algorithm in terms of noise resistance and initial guess. Over here, an initial learning rate is required to define the step size of the algorithm. Of course, some compromises are made to play with step size so that an equilibrium is found between the convergence speed and the local minimum. The concept of a learning algorithm, as the author puts it originates from the training of neural networks. Batch/incremental optimisation of pose graphs and loop closure are main focus of this work.

On a similar note, the work of, [Konolige & Agrawal 2008] discuss the issue of pose graph pruning using a precise motion estimation algorithm based on a non-linear least square (NLLS) approach. Continous visiting of the same places does not cause the map to bulge out, meaning loop closures are carried out quite efficiently. The interesting part of the results include the validation over large outdoor datasets. Furthermore, Konolige and Bowman, [Konolige & Bowman 2009] attempts to solve the problem of skeletal graph reduction using view clustering based on the ratio of inlier to oulier matches. Again, the concepts of lifelong mapping relies on factors discussed above such as batch/incremental mapping, loop closing as well as map repair to handle the problem of dynamic environments.

Johannsson *et al.* [Johannsson *et al.* 2013] further add that the size of the optimisation problem related to the pose graph should be constrained by the size of the explored environment and be independant upon the exploration time. Over here, a mobile robot operated over a multi-storey building where the floors were adventured using a lift in between transitions. The system along with visual odometry was equipped with an IMU and with wheel odometry which interfaces when the vision system is unable to estimate motion, for e.g, when textureless walls are traversed, prevailing lighting conditions and camera occlusion. The advantages are two-fold; the robot was kept running even at visual odometry failures which later helped in correcting the map in subsequent runs. Furthermore, a keyframe representation alone has its drawbacks - discarding intermediate frames lead to consequent loss of information while their approach adapted from an Exactly Sparse Extended Information Filter (ESEIF) maintain both sparseness and consistency.

[Kaess *et al.*] introduces a technique called "incremental Smoothing and Mapping

(iSAM) which is practical for large scale environments and supports online data association". The Information filter based SLAM is formulated as a Least Squares Problem (LSP) where the solution is based on matrix factorisation. The Information matrix is exploited for measurement updates and estimation uncertainties. Most recently, [McDonald *et al.* 2013] extended the work iSAM framework for robust long time visual mapping and navigation. The key problems tacked are the long term repeatability operations of a mobile robot with map update in real time. They employ anchor nodes (an alternative to weak links) to construct a pose graph. Experimentations are carried out at multiple excursions of the robot and at each session, a seperately associated pose graph is built.

The distinct pose graphs in [McDonald *et al.* 2013], are thereafter stiched together by inducing constraints upon succesive interconnected nodes pertaining to seperate sessions. Feature tracking is first initialised using a GPU-based KLT tracker where the point clouds from the left image are successfully matched to the right image. An initial rough pose is estimated using 3 -point RANSAC algorithm. Finally, a refined pose is obtained by iteratively minimising the reprojection errors using a classical Levenberg-Marquardt optimisation algorithm. The system fares relatively well when tested both outdoor and indoor but its caveats are: tracking failures at high speed motion and at textureless environments where feature initialisation becomes difficult. Nevertheless, promissing results are displayed which has enlarged the author's perspective for mapping at different timescales.

2.5 Conlusion

As examplified in literature, VO is becoming a solid component for egomotion estimation due to the richness of information it provides . Furthermore, the mathematical tools developped over the years has shown that solid concepts can be used to overcome the limitations of other expensive sensors (Lidars, RTK GPS). Two mainstream techniques of VO are feature and dense based. Whilst both of them have their relative pros and cons, however, these days, dense based technique is becoming the preferred choise due to the advantage of bypassing a prior feature extraction stage, thereby making use of the entire information content of the sensor. The latter eventually fails when the difference in viewpoint is so large that the amount of outliers greatly exceeds that of the inliers which causes optimisation techniques to break down. In this case, feature based methods complements this limitation by providing an initial raw estimation which can then be refined using dense technique. This procedure is normally used for localisation problems where an initial estimate of the pose is not available.

Various solutions have been proposed in order to provide an approximate model of the environment. Though occupancy grids remain an integral part for indoor exploration due to its ease of stochastic space discretisation, however, the solution is not viable for outdoor applications. Octomaps attempt to bridge this gap by providing a more compact representation but even though it is implemented in the most efficient way, the huge amount of information obtainable for ourdoor scenes remains a major hurdle, because at some point of

time, memory capacity is bound to saturate. Furthermore, discretising human made spaces are still achievable and within reach of implementation, but for vast scale environment, discretising over hundreds of kilometres is simply not appropriate. Other representations such as feature or topological maps provide an alternate and efficient way to model the environment, whereby, instead of storing the entire perceptual space, only pertinent components (landmarks) are stored. An interesting proposition is that of topometric maps which combines two representations in a single framework so as to maximise the benefits of each one of them. In this work, we make explicit use of this representation to model the environment in an efficient way whereby VO is used as the backbone of metric maps. Secondly, instead of storing all the incoming information of the sensor, careful selection of keyframes render this representation compact and sparse.

Finally, a sneak peak of lifelong mapping is given which is emerging out as the hot topic of mobile robotics these days. Though we do not explicitly treat this problem on our work, in an initial attempt of tackling the subject, a thorough treatment of stable and unstable points will be discussed in the second fold of this thesis. The cross road at which we summarise our work is a direct extrapolation to lifelong map learning.

CHAPTER 3

Projective Geometry and Spherical Vision

3.1 Introduction

Projective geometry serves as a mathematical framework for 3-D multiview imaging and 3-D computer graphics. It is used to model the image formation process, generate synthetic images, or reconstruct 3-D objects from multiple images. In order to upgrade the projective reconstruction into a metric one, 3-D vision is divided or stratified into four geometry groups; projective, affine, metric and Euclidean forming the basis of any 3-D reconstruction [Henrichsen 2000]. To model lines, planes or points in 3-D space, the *Euclidean geometry* is usually employed. However this geometric tool presents two major drawbacks; the difficulty of modelling points at infinity which can be viewed as two railway lines intersecting at infinity and the second one being the projection of a 3-D point onto an image plane which requires a perspective scaling operation. As a scale factor is a parameter, perspective scaling requires a division that becomes a non-linear operation [Morvan 2009]. Nevertheless, projective geometry presents an attactive framework such as *homogeneity* shadowing the above-mentioned disadvantages. This concept shall be elaborated in the first part of this chapter.

After a comprehensive introductory overview of the generic theory of stereo vision and projective geometry, in the second fold, the subject of spherical vision is introduced by using the concepts underlined in the previous part. The idea of spherical representation of a captured set of images comes from the fact that, for long, artists or people from the field of photography have had a strong interest in building up panoramic pictures to depict real or virtual scenes. Panoramic photography is a concept of cropping down images to a relatively wide aspect ratio. Later on, pertinent efforts from experts of the field of electronics came up with interesting products that have had considerable success in photography, cinematography, as well as the consumer market. For long, researchers from various field studied the behaviour or movements of animals and insects in their immediate environment. This active research area is known as ***ethology***. It has been found that these creatures use the advantage of a wide field of view in order to displace or locate themselves from their original motions. Only in the 90s, researchers from the field of computer vision and robotics have come on a common ground to study the means of implementing these kind of locomotions to mobile unmanned ground, aerial and subsea vehicles. The idea be-

hind is to conceive a compact model that can represent a maximum number of information of the environment. The greater the field of view (FOV), closer the model will be from a real environment, the better the navigation [Mei 2007, pan 2012]. Since then, efforts have been concentrated around building suitable sensors or devising methods to be able to capture a 360 degree view of the environment.

Recent developments in the area of spherical perspective projection include cameras with wide objective angle giving fisheye images, omnidirectional catadioptric cameras [Mei 2007] or multi-baseline cameras (to name a few), [Meilland *et al.* 2010, Meilland *et al.* 2011a]. The multitude of advantages it offers makes it attractive for navigation and localisation applications. We shall elaborate on the idea of generating spherical panoramic images from projective geometry. In order to reap the benefits of this particular configuration, several multi-camera systems have been developped in our lab throughout the years. A brief description of conception and calibration techniques for each one of them will be exposed, based on the methodology used by Meilland, to serve as foundation tools for the subsequent chapters of the manuscript.

3.2 The camera model

A camera is a mapping between the 3-D world and a 2-D image. Different models exist in practice but the most widely used models fall under the category of central projection which itself fissions into two major classes; camera models with a finite centre or models with centre taken to be at "infinity". In literature, the basic **pinhole model** is the basis of every exploitation describing the process of image formation. Since image formation is the result of a series of transformations of coordinates, a mathematical model under the assumptions of a pinhole camera model and Lambertian surfaces must account for three types of transformations:

1. coordinate transformations between the camera frame and the world frame;
2. projection of 3-D coordinates onto 2-D coordinates
3. coordinate transformation between possible choices of image coordinate frame.

The following sections will describe the intrinsic and extrinsic matrices which are the two main concatenating blocks making up the model.

3.2.1 Intrinsic parameters

Intrinsic parameters describe the internal parameters of a camera such as focal distance, radial lens parameters, image centre, skew factor. From figure 3.1, the centre of projection (the point at which all the rays intersect) is denoted as the *optical centre* or *camera centre* and the line perpendicular to the image plane passing through the optical centre as the

optical axis. Additionally, the intersection of the image plane with the optical axis is called the *principal point*.

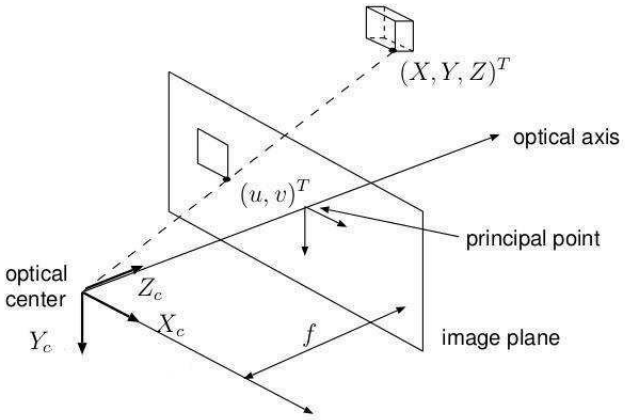

Figure 3.1: *Projection of a world coordinate onto the pin hole camera model plane.*

Consider a camera with the optical axis being collinear to the Z_c axis and the optical centre being located at the origin of a 3-D coordinate system. By similar triangles, the projection of a 3-D world point $(X, Y, Z)^T$ onto the image plane at pixel position $(u, v)^T$ can be written as:

$$u = \frac{Xf}{Z} \quad \text{and} \quad v = \frac{Yf}{Z}, \tag{3.1}$$

where, f denotes the focal length. To avoid such a non-linear division operation, the above equation can be converted into a linear form by the use of homogeneous transforms from the boon of projective geometry framework. Hence the above relation can be expressed in matrix notation by:

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \tag{3.2}$$

where, $\lambda = Z$ is the homogeneous scaling factor.

Since all image processing software tools identify each pixel location in pixel coordinates usually located at the top-left pixel of the image, it is necessary to transfer the coordinate system initially assumed to be at the centre of the camera coordinate system to the image coordinate system as shown in figure 3.2. In this process, the transformation from metric to pixel distance is also required as follows:

$$S_x = \frac{x}{u - u_0} \quad \text{mm/pixel}; \qquad S_y = \frac{y}{v - v_0} \quad \text{mm/pixel},$$

from where,

$$u = \frac{1}{S_x} x + u_0, \qquad v = \frac{1}{S_y} x + v_0 \tag{3.3}$$

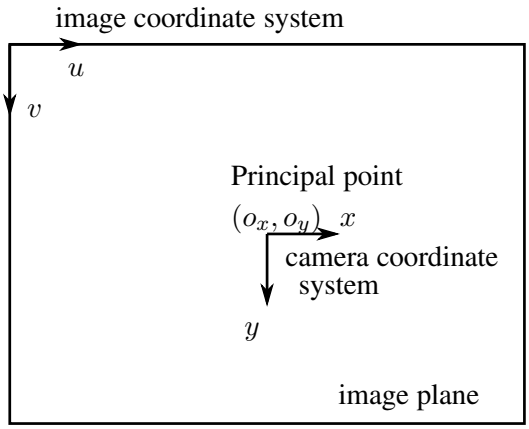

Figure 3.2: *Transformation from image to pixel frame.*

Equation 3.3 is again transformed into a homogeneous matrix and injected in equation 3.2 to give:

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{S_x} & \tau & u_0 \\ 0 & \frac{1}{S_y} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \tag{3.4}$$

where τ defines the skewness of the pixel which is assumed to be zero for recent digital cameras.

3.2.2 Extrinsic parameters

Extrinsic parameters indicate the external position and orientation of a camera in the 3-D world. Normally, any object in space is defined by its own coordinate frame. By setting up a fixed coordinate system known as the reference frame, the position and orientation of any point will be expressed in that frame. This is important to initialise any system in the euclidean space as rigid body motion is computed relative to an initial given position and orientation. Therefore, there arise the need to align any perceived object in space with respect to an initialised reference orthonormal frame which in our case is taken to be that of the camera. Any rigid point in space is subjected to two basic transformations; a translation defining the distance travelled and/or rotation defining the subjected twist. Figure 3.3 illustrates the idea described above.

A homogeneous matrix $\mathbf{T} \in \mathbb{SE}(3) \subset \mathbb{R}^{4\times4}$, belonging to the Euclidean group is defined as:

$$\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}, \tag{3.5}$$

where, $\mathbf{R} \in \mathbb{SO}(3) \subset \mathbb{R}^{3\times3}$ is a rotation matrix of the special orthogonal group and $\mathbf{t} \in \mathbb{R}^3$

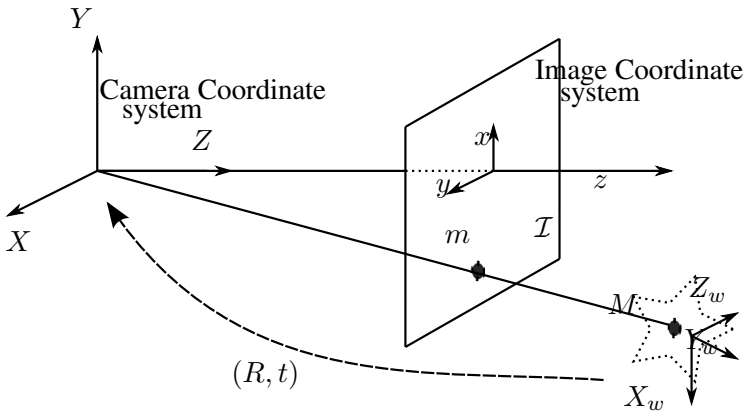

Figure 3.3: *Transformation from world coordinate frame to camera coordinate frame.*

denotes the translational vector. While an inverse transformation of 3.5 is given by:

$$\mathbf{T}^{-1} = \begin{bmatrix} \mathbf{R}^T & -\mathbf{R}^T\mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}. \tag{3.6}$$

3.2.3 Projection equation

The projection equation mapping a 2-D image point $\mathbf{p}$ to a 3-D world coordinate $\mathbf{P}$ is obtained by plugging equation 3.5 into 3.4 to obtain:

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{S_x} & \tau & u_0 \\ 0 & \frac{1}{S_y} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}. \tag{3.7}$$

In compact form, 3.7 can be written as:

$$\lambda\mathbf{p} = [\mathbf{K} \quad \mathbf{0_3}][\mathbf{R} \quad \mathbf{t}]\mathbf{P} \qquad \text{or} \qquad \begin{bmatrix} \lambda_i u_i \\ \lambda_i v_i \\ \lambda_i \end{bmatrix} = \begin{bmatrix} M_{11} & M_{12} & M_{13} & M_{14} \\ M_{21} & M_{22} & M_{23} & M_{24} \\ M_{31} & M_{32} & M_{33} & M_{34} \end{bmatrix} \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix}, \tag{3.8}$$

where $[\mathbf{M}]_{ij}$ is known as the projection matrix. Although the derivation has been tackled starting from 2-D to 3-D projection, the camera intrinsic and extrinsic parameters are usually unknown. The projection matrix interacts between the inner and the outer world of 3-D vision. In order to obtain the intrinsic and extrinsic parameters of the camera, the projection matrix is exploited and worked back. This technique shall be covered in section of calibration. However, an immediate parameter that can be recovered from the $[\mathbf{M}]_{ij}$ is the lens centre with respect to the world coordinates.

By expanding 3.8, the lens centre can be found as follows:

$$\lambda_i u_i = M_{11}X_i + M_{12}Y_i + M_{13}Z_i + M_{14} \tag{3.9a}$$
$$\lambda_i v_i = M_{21}X_i + M_{22}Y_i + M_{23}Z_i + M_{24} \tag{3.9b}$$
$$\lambda_i = M_{31}X_i + M_{32}Y_i + M_{33}Z_i + M_{34}, \tag{3.9c}$$

From 3.9, it can be easily distinguished that:

$$u_i = \frac{M_{11}X_i + M_{12}Y_i + M_{13}Z_i + M_{14}}{M_{31}X_i + M_{32}Y_i + M_{33}Z_i + M_{34}}$$
$$v_i = \frac{M_{21}X_i + M_{22}Y_i + M_{23}Z_i + M_{24}}{M_{31}X_i + M_{32}Y_i + M_{33}Z_i + M_{34}}$$

Given that a ray of light passing through any image point must also pass through the lens centre for a pinhole camera model, for all arbitrary image coordinates (u, v), the lens centre (X_c, Y_c, Z_c) is obtained by solving the following set of equations for a unique solution:

$$M_{11}X_c + M_{12}Y_c + M_{13}Z_c + M_{14} = 0 \tag{3.10a}$$
$$M_{21}X_c + M_{22}Y_c + M_{23}Z_c + M_{24} = 0 \tag{3.10b}$$
$$M_{31}X_c + M_{32}Y_c + M_{33}Z_c + M_{34} = 0 \tag{3.10c}$$

3.3 Calibration

Camera calibration involves the estimation of both extrinsic and intrinsic parameters. There are various techniques in literature to determine these unknowns. In particular, the direct parameter calibration method wherein an initial estimate of the principal point (figure 3.2) from the *orthocenter theorem* leads to an estimation of all the other parameters [Bebis 2012], [Sung 2008]. 3-D reference object based calibration is an efficient technique with good achievable precision but the approach require an expensive apparatus and elaborate setup. Self calibration is a method that bypasses the need of a calibration object. By moving the camera in a static scene, it turns out that the rigidity of the scene provides in general two constraints on the cameras' internal parameters, enough to recover both the internal and external parameters. However, this method is still a subject of research and is not mature. The methodology that will be outlined in the sequel is the one proposed by Zhang, [Zhang 2000] which provides an accurate, inexpensive technique with focus on desktop vision systems.

3.3.1 Calibration with a planar pattern:

A more commonly adopted approach consists of capturing several images of a known planar object, such a checkerboard pattern as shown in figure 3.4. This technique only requires the camera to observe a planar pattern from a few different orientations. Although the min-

imum number of orientations is two if pixels are square, four or more different orientations will result in better quality and increased robustness of the final result. The camera or the planar pattern can either be moved. The motion does not need to be known, but should not be a pure translation. The movements should sweep as much as possible the field of view (FOV) of camera with varying distances to obtain a good calibration set.

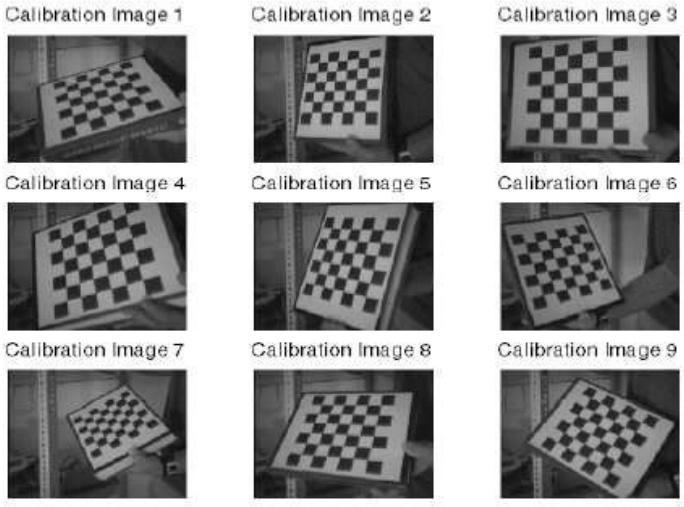

Figure 3.4: *Calibration with images of a checkerboard pattern.*

3.3.2 Computation of Homography Transform:

A homography is a transformation mapping a point from one 2-D (planar) plane to the other. To give a brief illustration, we consider a point $\mathbf{x_1}$ in a first image that is the image of some point, say p on the plane P. Then its corresponding second image $\mathbf{x_2}$ is uniquely determined as $\mathbf{x_2} \sim H\mathbf{x_1}$, where $\sim$ indicates an equality up to a scale factor. The equation is also referred to as a *planar homography* induced by a plane P, where H introduces a special mapping between points in the first image and those in the second one.

Since the world reference frame can be freely chosen, from figure 3.4, it is aligned to the board such that the world coordinate system is on $Z = 0$. Then from equation 3.8,

$$\lambda \mathbf{p} = \mathbf{K}\begin{bmatrix}\mathbf{r_1} & \mathbf{r_2} & \mathbf{r_3} & \mathbf{t}\end{bmatrix}\begin{bmatrix}X\\Y\\0\\1\end{bmatrix} = \mathbf{K}\begin{bmatrix}\mathbf{r_1} & \mathbf{r_2} & \mathbf{t}\end{bmatrix}\begin{bmatrix}X\\Y\\1\end{bmatrix}, \tag{3.11}$$

$$\implies \lambda \tilde{\mathbf{p}} = \mathbf{H}\tilde{\mathbf{P}}, \qquad \text{with} \quad \mathbf{H} = \mathbf{K}\begin{bmatrix}\mathbf{r_1} & \mathbf{r_2} & \mathbf{t}\end{bmatrix} \tag{3.12}$$

However, in practice, equation 3.12 does not hold because of noise in the extracted image points. Assuming that $\mathbf{p_i}$ is corrupted by Gaussian noise with zero mean and covariance

matrix $\mathbf{\Lambda_{p_i}}$, then the maximum likelihood function estimation of $\mathbf{H}$ is obtained by the minimisation of the following function:

$$\underset{H}{\operatorname{argmin}} \sum_i \|\mathbf{p_i} - \hat{\mathbf{p}}_\mathbf{i}\|^2 \tag{3.13}$$

where,

$$\hat{\mathbf{p}}_\mathbf{i} = \frac{\mathbf{1}}{\mathbf{h_3^T}} \begin{bmatrix} h_1^T M_i \\ h_2^T M_i \end{bmatrix}, \quad \text{with } \mathbf{h_i} \text{ being the } i\text{th row of } \mathbf{H}. \tag{3.14}$$

Equation 3.13 above, is a non-linear optimisation problem which can be solved using *Gauss Newton* or *Levenberg-Marquardt Algorithm*.
With $\mathbf{x} = [\mathbf{h_1^T} \quad \mathbf{h_2^T} \quad \mathbf{h_3^T}]^\mathbf{T}$, equation 3.12 can be written as follows:

$$\begin{bmatrix} \widetilde{P}^T & 0^T & -u\widetilde{P}^T \\ 0^T & \widetilde{P}^T & -v\widetilde{P}^T \end{bmatrix} \mathbf{x} = \mathbf{0} \tag{3.15}$$

For n points, n above equations are obtained which can be written in matrix form $\mathbf{Ax} = \mathbf{0}$, where $\mathbf{A}$ is a $2n \times 9$ matrix. The solution $\mathbf{x}$ is then extracted from the *Singular Value Decomposition* (SVD) of $\mathbf{A}$, where the eigenvector of $\mathbf{A^T A}$ associated to the smallest eigenvalues results in the components of $\mathbf{x}$.

3.3.3 Computation of the Calibration Matrix:

For each image, a homography can be estimated as described in the previous section. From $\mathbf{H} = [\mathbf{h_1} \quad \mathbf{h_2} \quad \mathbf{h_3}]$, equation 3.12 is rewritten as:

$$[\mathbf{h_1} \quad \mathbf{h_2} \quad \mathbf{h_3}] = \lambda \mathbf{K} [\mathbf{r_1} \quad \mathbf{r_2} \quad \mathbf{t}],$$

where λ is a scalar. Since vectors $\mathbf{r_1}$ and $\mathbf{r_2}$ are orthonormal (fundamental property of rotation matrices), the following two equations are obtained and give two constraints on the internal parameters of the camera:

$$\mathbf{h_1^T K^{-T} K^{-1} h_2} = \mathbf{0} \tag{3.16}$$

$$\mathbf{h_1^T K^{-T} K^{-1} h_1} = \mathbf{h_2^T K^{-T} K^{-1} h_2} \tag{3.17}$$

From equation 3.7, redefine $\mathbf{K}$ as:

$$K = \begin{bmatrix} \alpha & \gamma & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{3.18}$$

then,

$$\xi_{\mathbf{i}} = \mathbf{K^{-T}K^{-1}} \equiv \begin{bmatrix} \xi_1 & \xi_2 & \xi_4 \\ \xi_2 & \xi_3 & \xi_5 \\ \xi_4 & \xi_5 & \xi_6 \end{bmatrix} = \begin{bmatrix} \frac{1}{\alpha^2} & -\frac{\gamma}{\alpha^2\beta} & \frac{v_0\gamma - u_0\beta}{\alpha^2\beta} \\ -\frac{\gamma}{\alpha^2\beta} & -\frac{\gamma^2}{\alpha^2\beta^2} + \frac{1}{\beta^2} & -\frac{\gamma(v_0\gamma - u_0\beta)}{\alpha^2\beta^2} - \frac{v_0}{\beta^2} \\ \frac{v_0\gamma - u_0\beta}{\alpha^2\beta} & -\frac{\gamma(v_0\gamma - u_0\beta)}{\alpha^2\beta^2} - \frac{v_0}{\beta^2} & \frac{\gamma(v_0\gamma - u_0\beta)}{\alpha^2\beta^2} + \frac{v_0^2}{\beta^2} + 1 \end{bmatrix}$$

Noticing that $\boldsymbol{\xi}_i$ is a symmetric matrix, $\boldsymbol{\xi}_i = (\xi_1, \xi_2, \xi_3, \xi_4, \xi_5, \xi_6)^T$ and denoting the i^{th} column vector of $\mathbf{H}$ by $\mathbf{h_i} = [h_{i1}, h_{i2}, h_{i3}]$, the following equation is derived:

$$\mathbf{h_i^T \boldsymbol{\xi}_i h_j} = \mathbf{v_{ij}^T} \xi, \tag{3.19}$$

where,

$$\mathbf{v_{ij}} = [h_{i1}h_{j1}, h_{i1}h_{j2} + h_{i2}h_{j1}, h_{i2}h_{j2}, h_{i3}h_{j1} + h_{i1}h_{j3}, h_{i3}h_{j2} + h_{i2}h_{j3}, h_{i3}h_{i3}]^T.$$

From 3.16, two homogeneous equations can be deduced:

$$\begin{bmatrix} v_{12}^T \\ (v_{11} - v_{12})^T \end{bmatrix} \xi = 0 \tag{3.20}$$

For n images or n homographies, the above equation is stacked n times and the following is obtained:

$$\mathbf{V}\xi = \mathbf{0}, \tag{3.21}$$

where $\mathbf{V}$ is a $2n \times 6$ matrix. The general solution is then obtained again using SVD. Matrix $\boldsymbol{\xi}_i$ is defined up to a scalar $\boldsymbol{\xi}_i = \lambda \mathbf{K^{-T}K^{-1}}$, and it is then possible to extract the internal parameters of the camera , once the vector $\boldsymbol{\xi}_i$ is known as follows:

$$\begin{aligned} v_0 &= \frac{(\xi_2\xi_4 - \xi_1\xi_5)}{\xi_1\xi_3 - \xi_2^2} \\ \lambda &= \xi_6 - \frac{[\xi_4^2 + v_0(\xi_2\xi_4 - \xi_1\xi_5)]}{\xi_1} \\ \alpha &= \sqrt{\frac{\lambda}{\xi_1}} \\ \beta &= \sqrt{\frac{\lambda\xi_1}{(\xi_1\xi_3 - \xi_2^2)}} \\ \gamma &= -\frac{\xi_2\alpha^2\beta}{\lambda} \\ u_0 &= \frac{\gamma v_0}{\alpha} - \frac{\omega_4\alpha^2}{\lambda} \end{aligned}$$

The external parameters for each image can be also calculated from equation 3.12, once the calibration matrix $\boldsymbol{K}$ is estimated:

$$\begin{aligned} \mathbf{r}_1 &= \lambda K^{-1}\mathbf{h}_1 \\ \mathbf{r}_2 &= \lambda K^{-1}\mathbf{h}_2 \\ \mathbf{r_3} &= \mathbf{r_1} \times \mathbf{r_2} \\ \mathbf{t} &= \lambda K^{-T}\mathbf{h}_3 \end{aligned}$$

where the scalar $\lambda = \frac{1}{\|K^{-1}\mathbf{h}_1\|} = \frac{1}{\|K^{-1}\mathbf{h}_2\|}$.

3.4 Stereo Calibration

This section unveils the basic geometry that relates images of points to their 3-D positions. The key interest is to reconstruct the relative pose (position and orientation) of the cameras as well as the locations of the points in space from their projection onto the two images. This time, the set up consists of two cameras of different positions and orientations staring at the same 3-D point in space.The estimation of coordinates of a 3-D point $\mathbf{P}$ can be performed in two steps. Firstly, given a selected pixel $\mathbf{p_1}$ in the image $\mathcal{I}_1$, the position of the corresponding pixel $\mathbf{p_2}$ in image $\mathcal{I}_2$ is estimated. $\mathcal{I}_1$ and $\mathcal{I}_2$ are known as stereo pairs while $\mathbf{p_1}$ and $\mathbf{p_2}$ are called a *point-correspondence*, coming from the projection of the same point $\mathbf{P}$ on both images $\mathcal{I}_1$ and $\mathcal{I}_2$. Secondly, after confirming the correspondence within a certain accuracy, the depth information and hence, the associated 3-D point can be computed by *triangulation*, using the geometry of the two cameras. The relationship involving the relationship between the stereo cameras is known as *epipolar geometry*, [Morvan 2009], [Ma *et al.* 2004].

3.4.1 Epipolar geometry

The concept is better explained along with an illustration of figure 3.5 below. A 3-D point $\mathbf{P}$ is projected through the camera centres $\mathbf{C}_1$ and $\mathbf{C}_2$, lying on the same plane π, known as the *epipolar plane*. The points $\mathbf{p}_1$ and $\mathbf{p}_2$ are the projected image of $\mathbf{P}$ on the image planes $\mathcal{I}_1$ and $\mathcal{I}_2$ respectively. A line joining the two camera centres $\mathbf{C}_1$ and $\mathbf{C}_2$ crosses the plane $\mathcal{I}_1$ and $\mathcal{I}_2$ at points e_1 and e_2 respectively, known as the *epipoles*. The distance between $\mathbf{C}_1$ and $\mathbf{C}_2$ is termed as the *baseline*. The image planes encounter the epipolar plane π, at the lines of intersection known as the *epipolar lines*. For a particular set up, $\mathbf{C}_1$, e_1 and $\mathbf{C}_2$, e_2 is fix for any point $\mathbf{P}_i$ observed in the scene. The emergence of the epipolar line provides a fundamental constraint for the task of point-correspondances which limits the search of the equivalent of $\mathbf{p}_1$ in $\mathcal{I}_2$, instead of an exhaustive time costly search of the entire image. Finally, the epipolar geometry can be described using a 3×3 rank-2 matrix, known as the *fundamental or essential matrix* $\mathbf{E}$, which is defined by $l_{\mathbf{2}} = \mathbf{E}\mathbf{p_1}$.

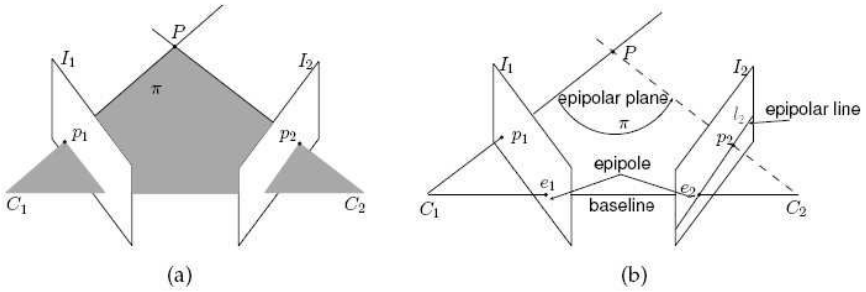

Figure 3.5: *Epipolar geometry.(a) The epipolar plane defined by the point* $\mathbf{P}$ *and camera centres* $\mathbf{C}_1$ *and* $\mathbf{C}_2$*. (b) Terminologies involved in epipolar geometry.*

3.4.2 Image Rectification

Image rectification is the process of transforming two images $\mathcal{I}_1$ and $\mathcal{I}_2$ such that their epipolar lines are horizontal and parallel. This procedure is particularly useful for depth-estimation algorithms because the search of point-correspondences can be performed along horizontal raster image lines. The technique consists of synthesising a common image plane $\mathcal{I}'$ and re-projecting the two images $\mathcal{I}_1$ and $\mathcal{I}_2$ onto this synthetic plane [Morvan 2009],[Fusiello *et al.* 2000], as shown in figure 3.6.

Consider a pixel $\mathbf{p}_1$ and its projections on the rectified image $\mathbf{p}'_1$ in $\mathcal{I}_1$ and $\mathcal{I}'_1$, respec-

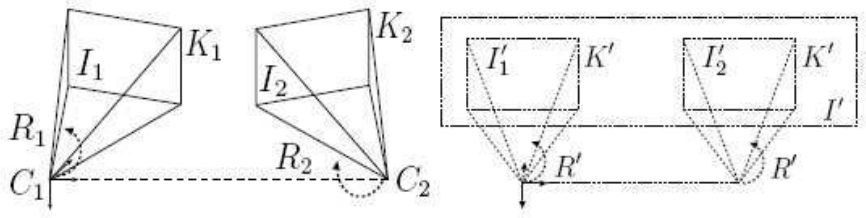

Figure 3.6: *Image rectification*

tively. Without loss of generality, it is assumed that the camera is located at the origin of the world coordinate system. The projection of a 3-D point $(X, Y, Z)^T$ onto the original and rectified images can be written as:

$$\lambda_1 \mathbf{p}_1 = \mathbf{K_1 R_1} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \quad \text{and} \quad \lambda'_1 \mathbf{p}'_1 = \mathbf{K' R'} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}, \tag{3.22}$$

with $\mathbf{R_1}, \mathbf{K_1}$ and $\mathbf{R}', \mathbf{K}'$ being the original and virtually rectified matrices respectively. Recombining the above equations leads to:

$$\frac{\lambda'_1}{\lambda_1} \mathbf{p}'_1 = \underbrace{\mathbf{K' R' R_1^{-1} K_1^{-1}}}_{\mathbf{H}_1} \mathbf{p}_1. \tag{3.23}$$

Similarly, the rectification of image $\mathcal{I}_2$ is obtained as follows:

$$\frac{\lambda_2'}{\lambda_2}\mathbf{p}_2' = \underbrace{\mathbf{K}'\mathbf{R}'\mathbf{R_2}^{-1}\mathbf{K_2}^{-1}}_{\mathbf{H}_2}\mathbf{p}_1. \tag{3.24}$$

3.4.2.1 Calculation of matrix $\mathbf{R}'$

The objective is to find a single rotation matrix $\mathbf{R}' = [\mathbf{r_1'} \quad \mathbf{r_2'} \quad \mathbf{r_3'}]^T$, common to the same axis of rotation of both cameras obtained as follows:

1. The row $\mathbf{r_1'}$ is defined parallel to the baseline going through the two camera centres $\mathbf{C}_1$ and $\mathbf{C}_2$, leading to $\mathbf{r_1'} = \frac{\mathbf{C_1}-\mathbf{C_2}}{\|\mathbf{C_1}-\mathbf{C_2}\|}$.
2. $\mathbf{r_2'}$ is chosen arbitrarily and in this case is taken to be the first row of $\mathbf{R}_1$. Hence $\mathbf{r_2'} = \mathbf{r_1} \times \mathbf{r_1'}$.
3. $\mathbf{r_3'}$ is defined orthogonal to $\mathbf{r_1'}$ and $\mathbf{r_2'}$ such that $\mathbf{r_3'} = \mathbf{r_1'} \times \mathbf{r_2'}$.
4. $\mathbf{r_k'}$ is normalised by $\mathbf{r_k'} := \frac{\mathbf{r_k'}}{\|\mathbf{r_k'}\|}, \quad k \in \{1, 2, 3\}$.

3.4.2.2 Point to Point correspondance

As explained earlier, the task is to find homologous points, *i.e.*, to the find the best match between a point c in the left image to a point $\mathbf{p}_r$ of the right image. Since it is very difficult to find a bijective mapping of $\mathbf{p}_l$ to $\mathbf{p}_r$, correlation is more practical using search windows of size $2W + 1$ for both images centered at $\mathbf{p}_l$ to $\mathbf{p}_r$. Some possible similarity measures are:

- Sum of Absolute Differences (SAD)
- Sum of Squared Differences (SSD)
- Normalised cross correlation (NCC)
- Zero centred Normalised cross correlation (ZNCC), which is an extrapolation of NCC

Even though the above-mentioned methods works well under certain conditions, they do have observe some shortcomings such as:

- Matching becomes difficult in textureless regions in low frequencies or constant grayscale values leading to an unreliable disparity estimate which is important in depth extraction.
- Some regions in a scene may not be visible from a selected viewpoint and hence correspondance cannot be achieved. This is known as the occlusion problem.

- Changes in contrast across the views. When capturing two images with different cameras, the contrast settings and illumination may differ. This results in different intensity levels across the views yielding unreliable matches.

3.4.2.3 Triangulation

The following step is to now extract the depth information after correspondance has been successfully achieved. In the case of a stereo rectified pair, the extrinsinc parameters is only a pure translation from the centre of right camera (baseline distance) to the one on the left if the latter is taken to be the reference as a usual rule of thumb. The depth is then extracted by:

$$Z = \frac{bf}{d}, \tag{3.25}$$

where b is the *baseline* distance , equivalent to a pure translation along x, denoted as t_x, f is the focal length computed from section 3.3.3, while d is the disparity between two corresponding points given by $d = p_l + p_r$. Consequently, a 3-D point $\mathbf{P} \in \mathbb{R}^3$, associated to a pixel $\mathbf{p}$ of an image is defined as:

$$\mathbf{P} = Z\mathbf{K^{-1}p} \tag{3.26}$$

3.4.2.4 Pose recovery

The final stage is now to find the extrinsic parameters linking two cameras forming a stereo pair. This relies on the theory of the *Essential matrix* introduced in section 3.4.1. A theorem based on pose recovery and the essential matrix, [Ma *et al.* 2004], states that *there exists exactly two relative poses* (R, T) *with* $R \in \mathbb{SO}(3)$ *and* $T \in \mathbb{R}^3$ *corresponding to a nonzero essential matrix* $E \in \mathcal{E}$, where $\mathcal{E}$ is referred to as the *essential space*.
From the *co-planarity constraint*, given 3, 3-D coplanar vectors **a,b,c**, their *triple product* is:

$$\mathbf{a}.(\mathbf{b} \times \mathbf{c}) = \mathbf{0}$$

Thus, for a stereo pair:

$$\begin{aligned}(\mathbf{R}\mathbf{p_r}).(\mathbf{t} \times \mathbf{p_l}) &= 0\\ \mathbf{p_r^T}\mathbf{R^T}(\mathbf{t} \times \mathbf{p_l}) &= 0\\ \mathbf{p_r^T}(\mathbf{R^T}\mathbf{T})\mathbf{p_l} &= 0\end{aligned}$$

$$\therefore \mathbf{E} = \mathbf{R^T}\mathbf{T}, \quad \mathbf{T} = [\mathbf{t}]_\times. \tag{3.27}$$

A major property of $\mathbf{E}$ is that it is of rank 2. This means that for each $\mathbf{p_r}, \mathbf{E}\mathbf{p}_l$ cannot generate more than two dimensions for all $\mathbf{p}_l$. For any two matched points, equation 3.27

can be explicitly written as:

$$\begin{bmatrix} x_r & y_r & 1 \end{bmatrix} \begin{bmatrix} E_{11} & E_{12} & E_{13} \\ E_{21} & E_{22} & E_{23} \\ E_{31} & E_{32} & E_{33} \end{bmatrix} \begin{bmatrix} x_l \\ y_l \\ 1 \end{bmatrix} = 0 \tag{3.28}$$

Next, the unknown elements of matrix **E** needs to be computed. For n point matches, expanding 3.28 and rearranging yields:

$$\begin{bmatrix} x_{l1}x_{r1} & y_{l1}x_{r1} & x_{r1} & x_{l1}y_{r1} & y_{l1}y_{r1} & y_{r1} & x_{l1} & y_{l1} \\ x_{l1}x_{r1} & y_{l1}x_{r1} & x_{r1} & x_{l1}y_{r1} & y_{l1}y_{r1} & y_{r1} & x_{l1} & y_{l1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{ln}x_{rn} & y_{ln}x_{rn} & x_{rn} & x_{ln}y_{rn} & y_{ln}y_{rn} & y_{rn} & x_{ln} & y_{ln} \end{bmatrix} \begin{bmatrix} E_{11} \\ E_{12} \\ E_{13} \\ E_{21} \\ E_{22} \\ E_{23} \\ E_{31} \\ E_{32} \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix}, \tag{3.29}$$

assuming that $E_{33} = 1$. The above equations can be solved by pseudo inverse or SVD to recover the essential matrix **E**. Once E is obtained, it is decomposed into its singular values as $\mathbf{E} = U\Sigma V^T$ where the pose (R, T) is recovered from two possibilities of R and T as follows:

$$R = UR_Z^T(\pm\frac{\pi}{2})V^T, \qquad T = UR_Z(\pm\frac{\pi}{2})\Sigma U^T. \tag{3.30}$$

3.5 Spherical Perspective projection

The choice of spherical perspective projection is partly motivated by the retina shapes often encountered in biological systems. For spherical projection [Ma *et al.* 2004], illustrated in figure 3.7, the image representation chosen to be the unit sphere, without loss of generality, $\mathbb{S}^2 = \{p \in \mathbb{R}^3 | \|X(p)\| = 1\}$.

Then, the spherical projection is defined by the map π_s from $\mathbb{R}^3$ to $\mathbb{S}^2$:

$$\pi_s : \mathbb{R}^3 \longrightarrow \mathbb{S}^2; X \longmapsto x = \frac{X}{\|X\|}. \tag{3.31}$$

Another fundamental reason, is the invariant property of spheres to rotation, meaning that, if the sphere is rotated around an axis ω at an angle θ, the projected point on the sphere remains unchanged. This property is important in rigid body transformations where image morphing is undesirable. However this projection has a main drawback related to the distribution of points on the sphere. Points around the poles are more clustered whereas those around the equator undergo a uniform distribution. A remedial way would be to analyse the sampling method used to define the sparsity of points on the sphere. As a matter of fact,

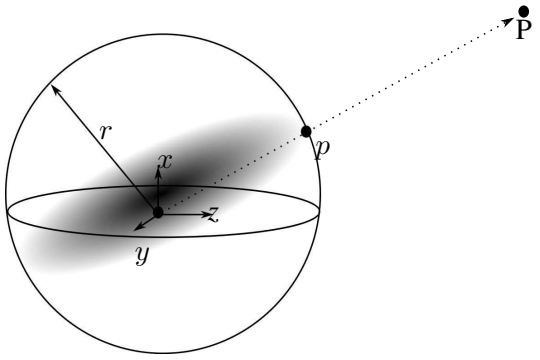

Figure 3.7: *Spherical perspective projection model; the image of a 3-D point p is the point x at the intersection of the ray going through the optical centre o of a sphere of radius r around the optical center. Typically* $r = 1$.

there exists different sampling techniques used, such as the Healpix, QuadCube, spiral, octahedral, to name a few.

3.6 Image Warping: Novel View Synthesis

Image warping, as defined in [Heckbert 1989], is the act of "distorting" a source image into a destination image according to a mapping between source space (u, v) and destination space (x, y). The mapping is usually specified by the functions $x(u, v)$ and $y(u, v)$.

Image warping is used in image processing primarily for the correction of geometric distortions introduced by imperfections imaging systems. Camera lenses sometimes introduce pincushion or barrel distortions, perspective views introduce a projective distortion, and other non linear optical components can create more complex distortions. In image processing, image warping is done typically to remove the distortions from an image, while in computer graphics, warping is usually introduced. Image warps are also used for the artistic purposes and special effects in interactive paint programs. For image processing applications, the mapping may be derived given a model of geometric distortions of a system, but more typically the mapping is inferred from a set of corresponding points in the source and destinations images. The point correspondence can be automatic, as for the stereo matching, or manual, as in paint programs. Most geometric correction systems support a limited set of mapping types, such as piecewise affine, bilinear, biquadratic, or bicubic mappings. Such mappings are usually parametrised by a grid of control points.

3.6.1 A formal description

Given an spherical image $\boldsymbol{\mathcal{I}}^* \in \mathbb{R}^{m \times n}$, related to a pixel intensity function $\boldsymbol{\mathcal{I}}^*(\mathbf{p}^*)$ and a reference frame $\boldsymbol{\mathcal{F}}^*$. The pixel coordinates of the image is defined as $\mathbf{p}^*$, with $u \in [0; m]$ and $v \in [0; n]$, and it is supposed that for each pixel $\mathbf{p}^*$, a corresponding depth information, $\boldsymbol{\mathcal{Z}} \in \mathbb{R}^+$ is known. A 3-D point in the Euclidean space in $\boldsymbol{\mathcal{F}}^*$, is denoted as $\boldsymbol{\mathcal{P}} = (\mathbf{p}^*, \boldsymbol{\mathcal{Z}}^*)$.

The set $\mathcal{S}^* = \{\mathcal{I}^*, \mathcal{Z}^*\}$ then defines an augmented spherical image embedding intensity as well as depth information.

Now, consider a second image $\mathcal{I}$ associated with a frame $\mathcal{F}$ and an intensity function $\mathcal{I}^w(\mathbf{p}^w)$ is the result of a transformation $\mathbf{T}(\tilde{\mathbf{x}})$ of the original image $\mathcal{I}^*$. The transformation subjected by the current image $\mathcal{I}$ in the current frame $\mathcal{F}$ is a 3-D displacement $\tilde{x} \in \mathbb{R}^6$ expressed in the reference frame $\mathcal{I}^*$, as illustrated in the figure 3.8. A warping function

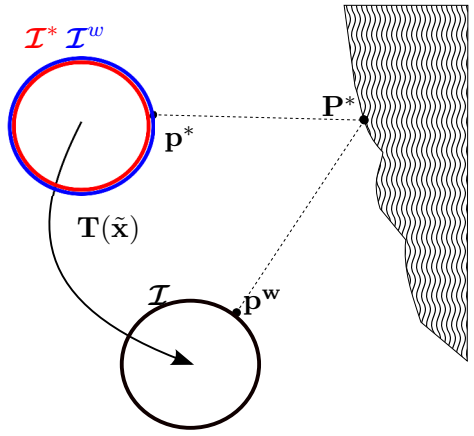

Figure 3.8: *A 3-D point $\mathcal{P}$ observed by the camera at instant t, projected on the reference frame as $\mathbf{p}^*$ while the same point observed by the camera at $t+1$, projected onto the current frame as $\mathbf{p}^w$. $\mathcal{I}^t_{t+1} = \boldsymbol{T(\tilde{x})}$, is the transformation mapping $\mathcal{I}$ onto $\mathcal{I}^*$,$\mathcal{I}^w$ being the warped image*

is now defined to combine the transformation described above. The projected point $\mathbf{p}^w$ is given by the following mapping:

$$\mathbf{p}^w = w(\mathbf{T}(\tilde{\mathbf{x}}); \mathcal{Z}; \mathbf{p}^*), \tag{3.32}$$

with its corresponding synthesised intensity value represented in $\mathcal{F}^*$ and under Lambertian assumption [Ma *et al.* 2004] is obtained by:

$$\mathcal{I}^w(\mathbf{p}^*) = \mathcal{I}\big(w(\mathbf{T}(\tilde{\mathbf{x}}); \mathcal{Z}; \mathbf{p}^*)\big) \tag{3.33}$$

Since direct point correspondences are not available ($\mathbf{p}^w \notin \mathbb{N}^2$), an interpolation function is normally required. The simplest non computation intensive interpolant can be a nearest neighbour or more smooth but involved functions like bilinear (4 nearest neighbours) or bicubic (16 nearest neighbours) [Keys 1981] interpolations.

3.7 Spherical Panorama

The idea of spherical panorama in a real time framework was lately developed by [Lovegrove & Davison 2010] where adjacent Keyframes are acquired by a purely rotating 3 degree of freedom (dof) camera. Keyframes as defined by the author are a set of

key historic camera poses associated with image data. With the help of a motion estimation module based on direct photometric image alignment, images are warped, projected and fused along successive viewpoints on a virtual sphere, tangential to the sensor by a mosaicing technique of [Szeliski 2006].

The above mentioned concept was further extrapolated by the works of [Meilland *et al.* 2010] [Meilland *et al.* 2011a] which dealt extensively with this kind of spherical representation. With the incoming of multiple images from perspective cameras, a novel synthesized high resolution (> 10 million pixels) 360^0 view of the environment is conceived as shown in figure (3.9). The total spherical projection equation of *N*, $\mathcal{I}_i$ intensity images transformed and fused on a unit sphere is defined by:

$$\mathcal{I}_s(\mathbf{q_s}) = \alpha_1 \mathcal{I}_1\big(w(\mathbf{K}_1, \mathbf{R}_1, \mathbf{q_s})\big) + \cdots\cdots + \alpha_N \mathcal{I}_N\big(w(\mathbf{K}_N, \mathbf{R}_N, \mathbf{q_s})\big), \tag{3.34}$$

where α_i is the coefficient of fusion, $\mathbf{K}_i$ is the intrinsic matrix parameters for each camera $\mathcal{C}_i$ as introduced in chapter (3) and $\mathbf{R}_i$ is the rotation matrix of $\mathcal{I}_i$ with respect to the sphere. A function $\bar{\mathbf{p}} = w(\mathbf{K}, \mathbf{R}, \mathbf{q_s})$ transfers a point of the unit sphere $\mathbf{q_s} \in \mathbb{S}^2$ in an image by the following perspective projection equation as illustrated in figure (3.10):

$$\bar{\mathbf{p}} = \frac{\mathbf{K}\mathbf{R}\mathbf{q_s}}{\mathbf{e}_3^T \mathbf{K}\mathbf{R}\mathbf{q_s}}, \tag{3.35}$$

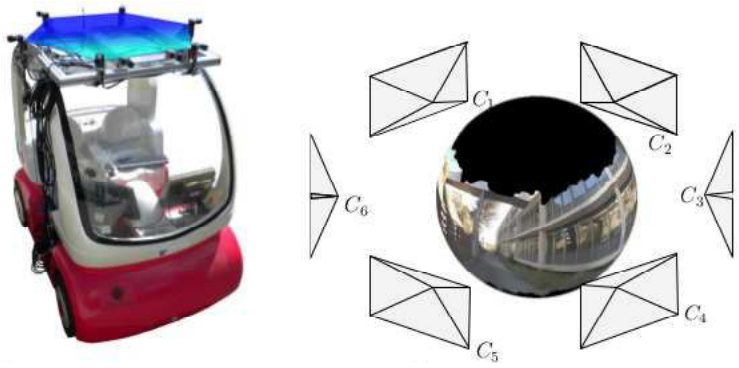

Figure 3.9: *Acquisition platform with multicamera system*

where $\mathbf{e}_3^T$ is the third vector component extractor of the denominator. The system provides a solution to a wide angle representation dedicated to urban environments. This design incorporates 6 cameras in a hexagonal configuration which purposely maintain a significant baseline between multiple divergent cameras. The augmented spherical panoramas provide both photometric and geometric (depth) information extracted between 6 stereo pairs using wide baseline dense matching. Moreover, the aspect of full view sensors maximises the observability of the environment and hence improves motion robustness by constraining all 6 dofs parametrised pose.

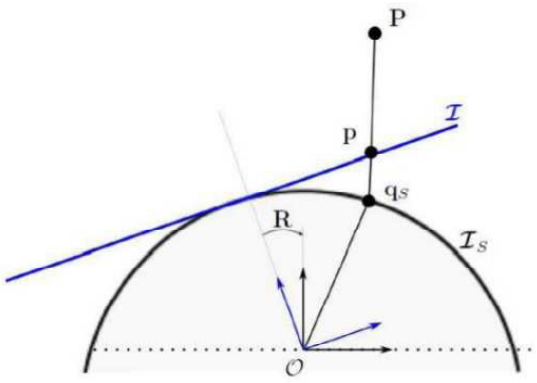

Figure 3.10: *Perspective image transformation on a sphere*

However, it's main caveat is that a unique centre of projection is assumed which is virtually set at the centre of gravity of the multicamera acquisition system. As a matter of fact, rotational motions (being independent of the scene geometry) are only considered while translational components are ignored. In cases where scene objects are close to the sensor, translation discrepancies between optical centres are non-negligible and hence the hypothesis of a unique centre of projection does not hold. In this case, artifacts related to parallax errors are visible on the reconstructed spherical panoramic images.

Figure 3.11: *Novel synthesised spherical panoramic image generated from the acquisition system illustrated in 3.9*

3.7.1 Novel System Design and Calibration

In order to bring the optical centres closer to a unique virtual centre of projection, the multi sensor configuration was re-designed maintaining the six cameras in a dual triangular-layer stereo configuration arrangement as shown in figure 3.12. To be able to construct an augmented sphere $\boldsymbol{\mathcal{S}} = \{\boldsymbol{\mathcal{I}}, \boldsymbol{\mathcal{D}}\}$, which consists of its photometric image along with its corresponding depth map, a rigorous calibration procedure is required in order to deduce the values of $\mathbf{R}$ and $\mathbf{K}$ as depicted in equation (3.34). Whilst intrinsic camera parameters are obtained from the technique described in chapter (3), overhere, an overview of the extraction of extrinsic parameter matrices and that of the depth map is argued for the sake

of completeness.

Figure 3.12: *Spherical RGBD outdoor acquisition system*

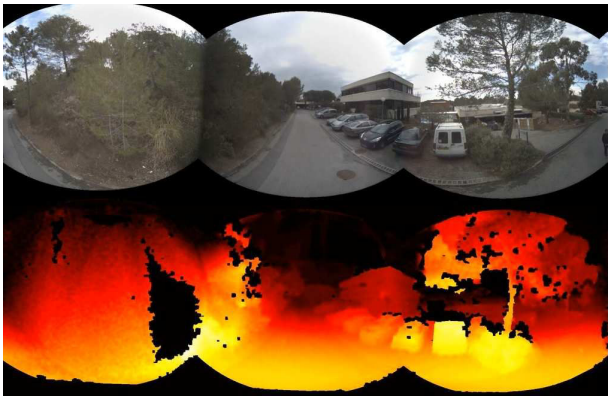

Figure 3.13: *Augmented RGBD sphere resulting from the acquisition system illustrated in figure 3.12*

For a such multi-baseline divergent camera system (120^0), only pairs of cameras observe the same parts of the scene which makes up several stereo pairs between the top and bottom camera layers. Further to the particularity of the system, the calibration methodology adopted follows a global loop closing approach applied independently to the top and bottom ring, yielding an optimization cost function in a bundle adjustment style. Given three intensity images $\mathcal{I}_1$, $\mathcal{I}_2$ and $\mathcal{I}_3$ of the same ring, the error minimization cost func-

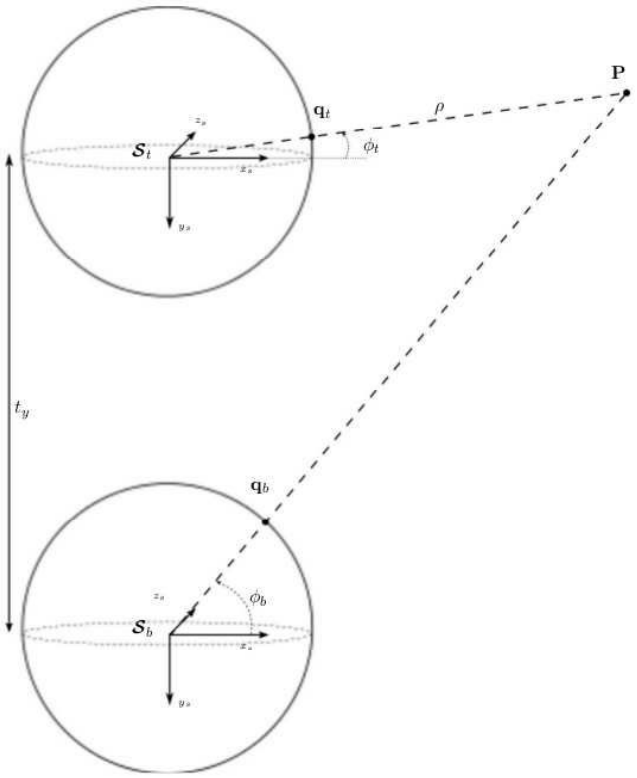

Figure 3.14: *Spherical Triangulation*

tion leads to:

$$\mathbf{e}^0 = \begin{bmatrix} \mathbf{e}_1 = \boldsymbol{\mathcal{I}}_2(w(\widehat{\mathbf{R}_2}\mathbf{R}(\mathbf{x}_2)\mathbf{K}_2), \mathbf{q_s})) - \boldsymbol{\mathcal{I}}_1(w(\mathbf{I}, \mathbf{K}_1), \mathbf{q_s}) \\ \mathbf{e}_2 = \boldsymbol{\mathcal{I}}_3(w(\widehat{\mathbf{R}_3}\mathbf{R}(\mathbf{x}_3)\mathbf{K}_3), \mathbf{q_s})) - \boldsymbol{\mathcal{I}}_1(w(\mathbf{I}, \mathbf{K}_1), \mathbf{q_s}) \\ \mathbf{e}_3 = \boldsymbol{\mathcal{I}}_3(w(\widehat{\mathbf{R}_3}\mathbf{R}(\mathbf{x}_3)\mathbf{K}_3), \mathbf{q_s})) - \boldsymbol{\mathcal{I}}_2(w(\widehat{\mathbf{R}_2}\mathbf{R}(\mathbf{x}_2)\mathbf{K}_2), \mathbf{q_s})) \end{bmatrix} \tag{3.36}$$

with $\boldsymbol{\mathcal{I}}_1$ fixed to identity, the motion parameters $\mathbf{x}_2$ and $\mathbf{x}_3$ are recovered using a classic Gauss-Newton unconstrained optimization algorithm. Explicit details about the error functions shall be elaborated in the subsequent chapter. Following a rectification process similar to that described in section (3.4.2) from which a rotation matrix $\mathbf{R}'$ common to both top and bottom spheres $\boldsymbol{\mathcal{S}}_t$ and $\boldsymbol{\mathcal{S}}_b$ is extracted, the fused and blended rectified sphere is then obtained as follows:

$$\boldsymbol{\mathcal{I}}_{\boldsymbol{\mathcal{S}}} = \alpha_1\boldsymbol{\mathcal{I}}_1(w(\mathbf{I}(\mathbf{R}')^\top), \mathbf{K}_1, \mathbf{q_s}) + \alpha_2\boldsymbol{\mathcal{I}}_2(w(\widehat{\mathbf{R}_2}(\mathbf{R}')^\top), \mathbf{K}_2, \mathbf{q_s}) + \alpha_3\boldsymbol{\mathcal{I}}_3(w(\widehat{\mathbf{R}_3}(\mathbf{R}')^\top), \mathbf{K}_3, \mathbf{q_s}), \tag{3.37}$$

where the fusion coefficients α_i are obtained from Laplacian Blending. Next, the disparity map is obtained from dense matching using techniques such as SAD block matching, *Efficient Large Scale Stereo Matching* (ELAS) [Geiger *et al.* 2010] or the *Semi Global Block Matching* (SGBM) [Hirschmuller 2006] followed by spherical triangulation (*cf.* figure (3.14)) of a world point $\boldsymbol{\mathcal{P}}$ projected as $\mathbf{q}_t = (\theta_t, \phi_t)$ on $\boldsymbol{\mathcal{S}}_t$ and $\mathbf{q}_b = (\theta_b, \phi_b)$ on $\boldsymbol{\mathcal{S}}_b$.

The disparity d_ϕ is written as:

$$d_\phi = \phi_b - \phi_t, \tag{3.38}$$

where distance $\rho \in \mathbb{R}^+$, associated to $\mathbf{q}_t$ is obtained by:

$$\rho = t_y \frac{\cos(\phi_t)}{\sin(d_\phi)}, \tag{3.39}$$

$t_y \in \mathbb{R}^+$ corresponds to the baseline between the rectified spheres. Consequently, world point $\boldsymbol{\mathcal{P}}$ can be reconstructed as follows:

$$\boldsymbol{\mathcal{P}} = \rho \begin{bmatrix} \sin(\theta_t)\cos(\phi_t) \\ \sin(\theta_t) \\ \cos(\theta_t)\cos(\phi_t) \end{bmatrix} \tag{3.40}$$

3.7.2 An innovative indoor Spherical RGBD sensor design

A new sensor for a large field of view RGBD image acquisition has been used in this work. This device integrates 8 Asus Xtion Pro Live (Asus XPL) sensors as shown in figure (3.15) and allows to build a spherical representation specifically for indoor applications.

The chosen configuration offers the advantage of creating full $360\,^\circ$ RGBD images of the scene isometrically, i.e. the same solid angle is assigned to each pixel. This permits to apply directly some operations, like point cloud reconstruction, photo consistency alignment or image subsampling. To build the images, the sphere $\mathbb{S}^2$ is sampled according to the resolution of our device, so that the longitude (θ direction) contains 1920 samples in the range $[0, 2\pi]$, while the latitude (ϕ direction) is sampled with the same spacing, containing 960 samples in the range $[-\pi/2, \pi/2]$. Since full range in ϕ is not observed by the sensor, only the useful range is stored which corresponds to a vertical FOV of $63\,^\circ$. The total resolution in pixels is 1920×640.

Figure 3.15: *Multi RGBD indoor acquisition system comprising of 8 Asus Xtion Pro live sensors*

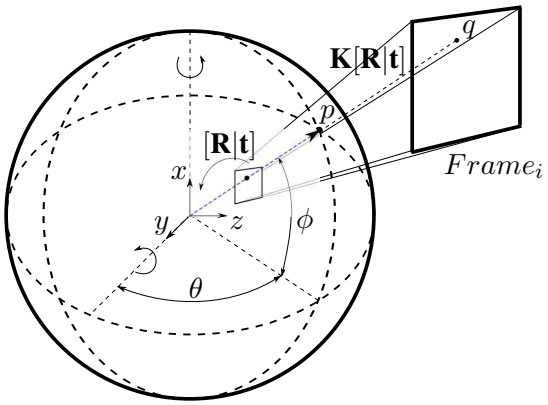

Figure 3.16: *Spherical RGBD construction making up our augmented sphere, $\boldsymbol{\mathcal{S}}$*

For spherical warping, a virtual sphere with the above sampling and radius $\rho = 1$ is

used to project the sample points into image coordinates (u, v), (*cf.* figure (3.16)). For that, the extrinsic calibration of each sensor is taken into account. Thus, a point p in $\mathbb{S}^2$ is parameterized in $\mathbb{R}^3$, using equation 3.40 as above, where ρ in this case equals 1 (unit sphere). The point $q = (u, v)$ on image coordinates is found by applying perspective projection to p, through the homogeneous matrix $\mathbf{M} = \mathbf{K}[\mathbf{R}|\mathbf{t}]$ where $\mathbf{K} \in \mathbb{R}^{3\times3}$ is the camera projection model and $[\mathbf{R}|\mathbf{t}] \in \mathbb{SE}(3)$ is the relative position of the camera (extrinsic calibration). Nearest neighbor interpolation is used to assign the intensity and depth values to the respective spherical images. Figure (4.18) depicts the panoramic views obtained from this spherical configuration.

Figure 3.17: *Example of spherical panoramic views obtained from our Office dataset using the multicamera system of figure (3.15)*

For the system described above, extrinsic calibration of the range cameras cannot be done by the calibration technique outlined in section (3.7.1) due to the non overlapping of frames yielding no point correspondences. Hence the approach adopted in [Fernández-Moral *et al.* 2014] provides a solution for such a system by making the most of structured environment geometries – observed planes from walls, ceilings, floors and other planar structures as shown in the set-up of figure (3.19). The approached methodology using this multi-camera system will be briefly discussed since it is part of the inner core of most of our experimental evaluations in this thesis.

Planar patches from the acquired depth images are segmented using a region growing technique and normals are computed from a set of points coming from the observed planes with their normal vectors $\mathbf{n}_i$ constrained to $\|\mathbf{n}_i\| = 1$. Consequently, the optimal distances d_i^* and covariance matrices Σ_i^* are extracted. Interframe plane correspondences are established by first providing heuristic geometrical constraints such as:

- the angle between normal vectors
- distances of both planes to camera centre
- threshold on the number of inter region pixels with respect to image pixels

Then the estimates are further refined using a RANSAC mechanism whilst monitoring the observability condition evaluated by the rank of the Fischer Information Matrix (FIM) as

follows:

$$\text{rank}(\sum_{i=1}^{N} \mathbf{n}_i \mathbf{n}_i^\top) = 3 \tag{3.41}$$

The ratio of the largest to the smallest eigenvalues of FIM gives a further indication of the distribution of planes in space such that a value of 1 states that planes are equally distributed in 3D space while a value tending towards 0 gives an ill-conditioned plane distribution in space.

The rotation and translational components of inter-sensor rigid transformation are then decoupled into two separate components encapsulated in a maximum log likelihood estimation problem leading to the following error functions:

$$\underset{\mathbf{x}=\mathbf{x}_2 \cdots \mathbf{x}_M}{\text{argmin}} \mathfrak{F}(\mathbf{x}) = \sum_{j=1}^{M} \sum_{k=j+1}^{M} \sum_{i=1}^{N} \lambda_i(j,k) \omega_i(j,k) \|\mathbf{R}(\mathbf{x}_j) \widehat{\mathbf{R}_j} \mathbf{n}_i^j - \mathbf{R}(\mathbf{x}_k) \widehat{\mathbf{R}_k} \mathbf{n}_i^k\|^2, \tag{3.42}$$

where,

$$\lambda_i(j,k) = \begin{cases} 1, & \textit{plane i observed by sensors j and k} \\ 0, & \textit{otherwise} \end{cases} \tag{3.43}$$

j and k are the indices of the M sensors, $\mathbf{n}_i^j$ and $\mathbf{n}_i^k$ are the normal vectors of the i^{th} plane observed from sensors j and k respectively. $\omega_i(j,k)$ is a weighting function based on Σ_i^*. In a similar formulation, the translational component resolves to:

$$\underset{\mathbf{x_t}=\mathbf{x_{t2}} \cdots \mathbf{x_{t}}_M}{\text{argmin}} \mathfrak{G}(\mathbf{x}_t) = \sum_{j=1}^{M} \sum_{k=j+1}^{M} \sum_{i=1}^{N} \lambda_i(j,k) \omega_i(j,k) \|d_i^j - d_i^k - \mathbf{t}(\mathbf{x}_t^j) \mathbf{R}_j \mathbf{n}_i^j + \mathbf{t}(\mathbf{x}_t^k) \mathbf{R}_k \mathbf{n}_i^k\|^2, \tag{3.44}$$

3.8 Conclusion

This chapter gives and overview the basic concepts of projective geometry and stereo vision extending to spherical vision. Starting from a pin hole camera model, an example of how an object in a world coordinate system is projected onto the camera frame is worked out using the intrinsic and extrinsic parameters of the camera. In order to compute the unknown parameters of the projection matrix, a calibration technique is devised using a checkerboard pattern. The method of [Zhang 2000] is highlighted from literature which is based on planar homography. These days, a plethora of software tool are available for camera calibration as listed in [Fraundorfer & Scaramuzza 2012].

Once the projective model is established, the next step is to recover depth information from stereo vision. A 2D-2D dense correspondence technique is used whereby world points are projecting on two camera frames at different viewpoints are associated using images

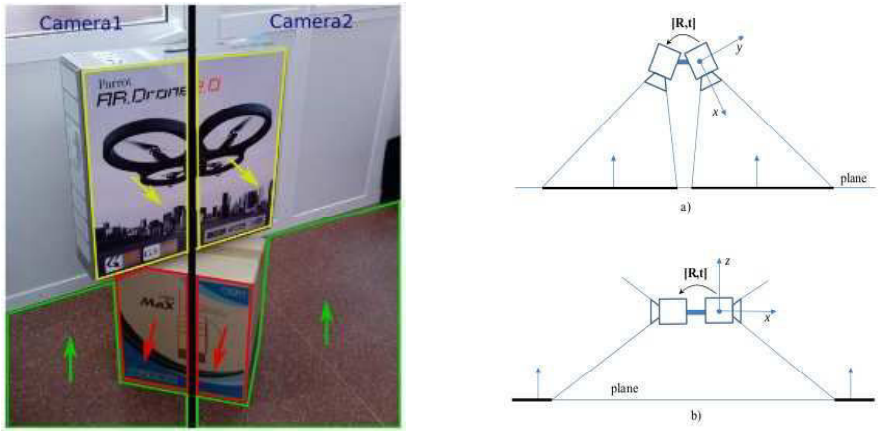

Figure 3.18: *left: Experimental set up for the calibration of two non-overlapping cameras(Adapted from [Fernández-Moral et al. 2014]), Right: Different sensor configurations for a stereo camera pair with (a) showing an adjacent set-up while (b) shows two opposite cameras observing the same plane*

Figure 3.19: *Top to bottom: photometric and geometric maps obtained from the device of figure (3.15) with their corresponding point cloud obtained by the spherical projection described in section (3.7.2)*

features bounded by the epipolar line. These image features are then back-projected in the world and triangulated using simple geometry to extract the depth information. The concept is referred to as epipolar geometry. Over here as well, a handful of software packages are available such as the *Efficient Large Scale Stereo Matching* (ELAS) [Geiger *et al.* 2010] or the *Semi Global Block Matching* (SGBM) [Hirschmuller 2006]. However, these feature based techniques are sensible to illumination changes which result in false depth estimates. To obtain consistent depth maps, further post treatments are required such as regularization techniques treated in [Newcombe 2012], but are more greedy in terms of computation resources. Though calibration techniques are not the main subject matter of this thesis, the handful of concepts involved help the reader to understand the most basic theory of 3D vision and how to infer the world geometry from solely images.

This chapter introduces the concept of wide FOV panoramic images with focus on spherical representations due to the multitude of advantages they offer. Spherical representation is invariant to rotation, hence to the orientation of the sensor. Furthermore, high resolution spherical images provide an enriched information content of the environment, highly desirable for localisation purposes. To harness the benefits of such systems, various sensors have been developed by our research team over the years concentrating on the conception of vision based navigation systems inclined towards problem solving under the SLAM framework.

In this context, the activities were centred around developing spherical outdoor and indoor multi-sensor systems. In an early development phase, 6 cameras arranged in hexagonal configuration was designed to conceive a wide 360^0 FOV spherical representation. The loophole of such a system is that each camera has its own centre of projection and the assumption of a unique centre of projection does not hold good, which poses problems for dense correspondence algorithms such as SAD Block Matching, SGBM or ELAS for depth extraction. To bridge the offset between the various optical centres, a new design has been conceived whereby this time, the 6 cameras are now spread along two layers in a triangular configuration. Our outdoor system outputs augmented spherical images consisting of both geometric and photometric information.

For indoor applications, a novel multiview short baseline camera system has been designed using 8 RGB-D Asus Xtion Pro live sensors to cover a 360^0 FOV. Initially destined for the gaming industry, these "Kinect-style" sensors provide both RGB and disparity images simultaneously, attracting the interest of robotics hobbyists for providing VSLAM solutions at very low cost. The sensor system operates in real-time with the construction of augmented spheres occurring online during the acquisition phase. A calibration technique based on plane to plane correspondences has been implemented due to the non-overlapping aspect of RGB-D images. It is observed that calibration errors undeniably affect the spherical mosaicing process, resulting in slight misalignment of the images. This introduces systematic errors inducing bias in the measurement estimate. This indoor sensor will be exploited at length in the validation of the various algorithms developed in this thesis.

CHAPTER 4

Spherical RGB-D Odometry

4.1 Introduction

This chapter is focussed on Visual Odometry (VO), defined as the process of estimating the relative motion of a mobile agent using vision sensors. This incremental technique computes the pose of a vehicle based on the movements induced by onboard cameras. Over the years, VO has been useful to compensate other similar techniques such as wheel odometry which is highly affected by dead reckoning in uneven terrains. On the other hand, global positioning system (GPS) has shown its limitation in aerial, underwater applications [Fraundorfer & Scaramuzza 2012] as well as in urban canyons. Current trends these days lean towards building photo-realistic 3D models with accurate geometry. Applications are vast and inexhaustive [Zhao *et al.* 2005]; 3D modelling of urban environments where 3D geometry and photometric information of the real world extending to city scales are recorded. Virtual reality applications, too represent a substantial potential, namely in the entertainment/ mobile applications' sector where realistic synthetic views of existing scenes are created out of few still images captured from consumer camera products.

Odometry techniques in general require accurate relative motion estimation to reduce trajectory drift. VO, which relies heavily on image contents requires at first hand good quality feature matching which makes the problem difficult [Fitzgibon 2003]. An important step prior to registration requires that data coming from two viewpoints should be put in correspondence. Two main approaches are identified; one which goes through an initial feature identification phase between the two data samples while the other uses dense correspondence technique [Fraundorfer & Scaramuzza 2012]. Over the last decade, VO coupled with SLAM approaches have evolved in two main categories; feature based and dense techniques. Feature based methods rely on a preceeding identification and extraction phase. Registration allows images which are further apart, but are affected by outliers. Dense technique, which uses the entire information content has become increasingly popular recently as registration is performed using a direct image alignment [Meilland & Comport 2013b]. The latter is generally more accurate but is restricted by smaller interframe displacements.

This chapter is decomposed as follows; starting from an initial optical flow model from literature, the 3D scene flow model is derived with application to Lukas-Kanade's direct image registration technique. This is further extended to our spherical RGB-D registration technique based on a first front on the photometric information. A second registration

technique based only on geometry is introduced inspired on the classical Iterative Closest Point (ICP) algorithm. To tackle the shortcomings of either cost function, a formulation is devised where both are incorporated in a single minimisation cost function. A pose graph representation is chosen as our mapping framework which consists of nodes and edges built on the backbone of VO. Each node is represented by a keyframe which stores the content of our augmented sphere. To deduce an optimal number of keyframes covering the explored environment, a criteria is generally required which gives an indication of the amount of changes which has occurred between two viewpoint changes. This criteria is of utmost importance since it helps in the reduction of data redundancy as well as suppression of tracking drift resulting from frame to keyframe registration. Two different criteria are highlighted, the first one based on the photometric cost function only while the second one is an abstraction of the VO pose uncertainty. A results section demonstrates the strengths and weaknesses of our proposed approach before wrapping up with a conclusion section.

4.2 From 2D Optic Flow to 3D Scene Flow

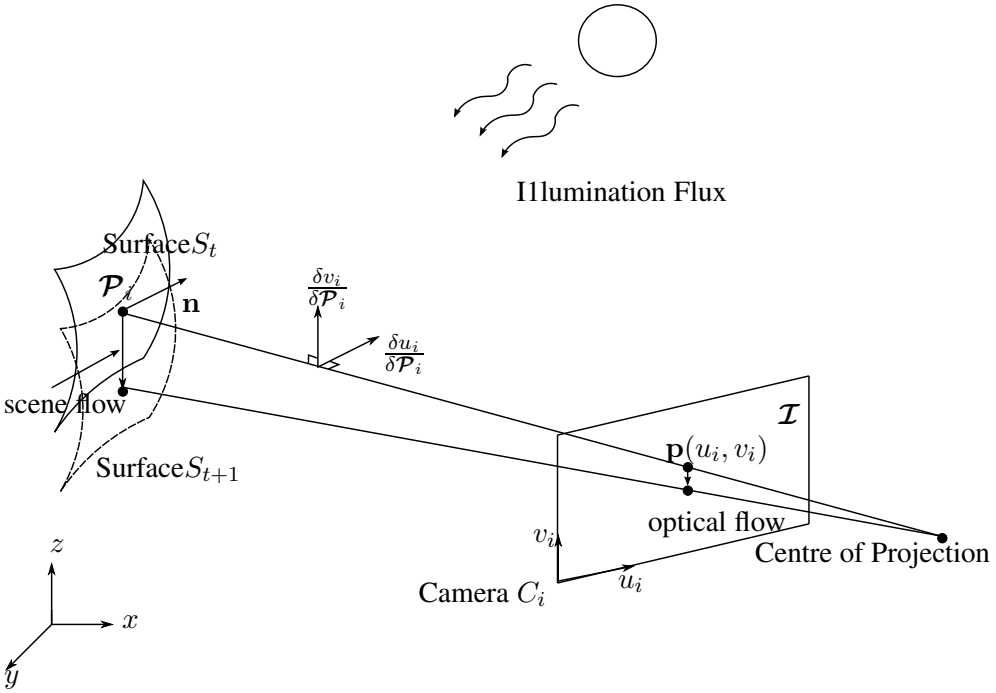

Figure 4.1: *Scene flow to optic flow model*

In previous chapters, image formation model has been studied to reconcile the idea of how objects in real world, making up millions of 3D points project on the camera frame. But a camera is a device measuring light intensities and not geometric primitives such as points, lines , edges for example. The idea of geometry inference out of image photometric measurements is rooted from the theory of optic flow devised by [Lucas & Kanade 1981].

Consider the scenario depicted in figure (4.1), where a non rigid surface S_t is moving with respect to a fixed coordinate system $f = (x, y, z)^T$. Given a motion perpendicular to

the normal $\mathbf{n}$ of S, the temporal surface transition of S^t to S^{t+1} relates to the scene flow as the instantaneous 3D motion of every point in the scene associated to S. The 3D motion of a world point $\boldsymbol{\mathcal{P}}$ projected back on the camera frame then defines the 2D optic flow of the scene into an image. One underlying assumption regarding the surface flow is that the illumination flux is constant throughout the motion and hence obeys the Lambertian hypothesis.

Let $\mathbf{p}_1$ be a pixel of an instantaneous frame $\boldsymbol{\mathcal{F}}_t$ and $\mathbf{p}_2$, it's new position at $\boldsymbol{\mathcal{F}}_{t+1}$ having undergone a certain random motion. The Brightness Change Constraint equation (BCCE) defined formally in [Harville *et al.* 1999] is given as:

$$\boldsymbol{\mathcal{I}}(u, v, t) = \boldsymbol{\mathcal{I}}(u + v_u(u, v, t), v + v_v(u, v, t), t + 1), \tag{4.1}$$

under the assumption that intensities undergo only local translation from one frame to the other in an image sequence. Phenomena such as occlusions, disocclusions, intensity variations are ignored in this formulation. $\boldsymbol{\mathcal{I}}(u, v, t)$ is the image intensity, $v_u(u, v, t)$ and $v_v(u, v, t)$ are the 2D components of the image velocity motion vector. Applying a first order Taylor series to the right hand side of equation (4.1) leads to:

$$\boldsymbol{\mathcal{I}}(u, v, t) = \boldsymbol{\mathcal{I}}(u, v, t) + \boldsymbol{\mathcal{I}}_u(u, v, t)v_u(u, v, t) + \boldsymbol{\mathcal{I}}_v(u, v, t)v_v(u, v, t) + \boldsymbol{\mathcal{I}}_t(u, v, t) \tag{4.2}$$

In compact form, the optical flow equation can be written as the following differential equation:

$$\frac{\partial \boldsymbol{\mathcal{I}}}{\partial \boldsymbol{\mathcal{I}}_u}\frac{d\boldsymbol{\mathcal{I}}_u}{dt} + \frac{\partial \boldsymbol{\mathcal{I}}}{\partial \boldsymbol{\mathcal{I}}_v}\frac{d\boldsymbol{\mathcal{I}}_v}{dt} + \frac{\partial \boldsymbol{\mathcal{I}}}{\partial t} = 0 \tag{4.3}$$

The optical flow equation (4.3) is in its most generic form and can be further extrapolated to encode the 3D motion of the pixel $\mathbf{p}$ given a certain projection model; be it perspective or spherical as discussed in chapter (3) for example. Moreover, the equation captures the relationship between the image velocity $\mathcal{V} = \begin{bmatrix} \frac{d\boldsymbol{\mathcal{I}}_u}{dt} & \frac{d\boldsymbol{\mathcal{I}}_v}{dt} \end{bmatrix} \in \mathbb{R}^2$ of $\mathbf{p}$ with its spatial and temporal derivatives $\nabla\boldsymbol{\mathcal{I}}$, $\boldsymbol{\mathcal{I}}_t$, directly measurable from images. A notable difference between optical flow and feature tracking is that in optical flow, focus is made onto one image location $\bar{\mathbf{p}}$ and the particles flow through $\bar{\mathbf{p}}$ is computed whereas in feature tracking, the particle $\mathbf{p}(t)$ is analysed instead and its location as it moves through the image domain is rather tracked. For multiple image measurements, equation (4.3) is cast in a linear model and solved for $\mathcal{V}$ using classic least means square [Harville *et al.* 1999]. The model presented in this section is also known as the **Normal Flow Constraint (NFC)** as defined in [Vedula *et al.* 2005].

4.2.1 Direct Photometric Registration

In this section, we describe the widely used Lucas-Kanade (L-K)[Lucas & Kanade 1981] formulation of direct image registration technique. It's goal is to iteratively align an image template $\boldsymbol{\mathcal{I}}^*(\mathbf{p}^*) \in \mathbb{R}^{m \times n}$ (taken as a reference), to an input image $\boldsymbol{\mathcal{I}}(\mathbf{p}) \in \mathbb{R}^{m \times n}$ using an objective function based on a sum of squared differences (SSD) similarity measure as follows:

$$\underset{\mathbf{x} \in \Re^n}{\operatorname{argmin}} \mathfrak{F}(\mathbf{x}) = \sum_{\mathbf{p} \in \mathbb{R}^{m \times n}} \left(\boldsymbol{\mathcal{I}}(w(\mathbf{p}; \mathbf{x})) - \boldsymbol{\mathcal{I}}^*(\mathbf{p}^*) \right)^2, \tag{4.4}$$

where $\boldsymbol{\mathcal{I}}(w(\mathbf{p}; \mathbf{x}))$ is the image warp described in section (3.6.1) which requires an image interpolation at sub-pixel location to obtain correspondences. Given an initial estimate of a hypothesised motion $\mathbf{d}$, L-K algorithm solves equation (4.4) for small increments of $\mathbf{d}$, *i.e.* $\triangle$ $\mathbf{d}$. The underlying motion $\mathbf{d}$ can be parametrised by a simple 2-D translation to more degrees of freedom (DOF) transformations such as euclidean (6 DOF), similarity (7 DOF), affine (12 DOF) for example [Hartley & Zisserman 2003](pg 73). Next, approaches to obtain an optimal solution of the cost function (4.4) are further elaborated.

4.2.1.1 Optimisation Tools

Equation (4.4) above is generally non linear in $\mathbf{x}$ and therefore, if a closed form solution is desired, it needs to be linearised. Assuming that the incremental pose of the camera is very small in time, equation 4.59 can be linearised with a Taylor series expansion around the neighbourhood of $\mathbf{x} = 0$, where $\mathbf{x} = [\boldsymbol{\omega}, \boldsymbol{\upsilon}] \in \mathbb{R}^6, \forall \boldsymbol{\omega}, \boldsymbol{\upsilon} \in \mathbb{R}^3$ are the angular and translational velocities (detailed in section 4.2.2). The problem is identified as a non-linear Least Means Square (LMS) unconstrained optimisation of the form:

$$\underset{x \in \Re^n}{\operatorname{argmin}} \mathfrak{F}(x) = \frac{1}{2} \sum_{i=1}^{mn} (e_i(x))^2 \equiv \frac{1}{2} \mathbf{e}^\top(\mathbf{x}) \mathbf{e}(\mathbf{x}) \equiv \frac{1}{2} \|\mathbf{e}(\mathbf{x})\|^2 \tag{4.5}$$

Assuming that the cost function f is differentiable and smooth so that the *Taylor expansion* is valid. The series expansion of a vector valued function **e(x)** about a point x_0 to the second order in **x** where $\triangle\, x = x - x_0$ is given by:

$$\mathbf{e(x)} = \mathbf{e}(x - x_0) = e(x_0) + \frac{\partial e(x_0)}{\partial x} \triangle x + \frac{1}{2} \triangle x^\top \frac{\partial^2 e(x_0)}{\partial x^2} \triangle x + \text{h.o.t} \tag{4.6}$$

Linearising at **x=0** and written in compact form, equation 4.6 leads to:

$$\mathbf{e}(\mathbf{x}) = \mathbf{e}(\mathbf{0}) + \mathbf{J}(0) \triangle \mathbf{x} + \frac{1}{2} \mathbf{M}(0, \mathbf{x}) \triangle \mathbf{x} + \mathbf{O}(\|\mathbf{x}\|^3) \tag{4.7}$$

From equations 4.5 and 4.7 the objective function to be minimised is written as:

$$\mathfrak{F}(\mathbf{x}) = \frac{1}{2}\|\mathbf{e}(\mathbf{0}) + \mathbf{J}(0) \triangle \mathbf{x} + \frac{1}{2}\mathbf{M}(0,\mathbf{x}) \triangle \mathbf{x}\|^2, \tag{4.8}$$

where the factor $\frac{1}{2}$ is induced without loss of generality,for mathematical convenience only and has no effect on the optimal solution x^* at $\nabla_{\mathbf{x}}\mathfrak{F}(\mathbf{x})|_{x=\tilde{x}} = 0$. $\mathbf{J}(.)$ and $\mathbf{M}(.)$ are the Jacobian and Hessian matrices respectively, both of dimensions $mn \times 6$.

The derivative of the cost function can be written as:

$$\nabla_{\mathbf{x}}\mathfrak{F}(\triangle \mathbf{x}) = (\mathbf{J}(0) + \frac{1}{2}\mathbf{M}(0,\mathbf{x}))^\top(\mathbf{e}(\mathbf{0}) + \mathbf{J}(0) \triangle \mathbf{x} + \frac{1}{2}\mathbf{M}(0,\mathbf{x}) \triangle \mathbf{x}), \tag{4.9}$$

where, the least square incremental update according to the Newton method resolves to:

$$\triangle \mathbf{x} = -\mathbf{Q}^{-1}\mathbf{J}(0)^\top\mathbf{e}(0), \tag{4.10}$$

where,

$$\mathbf{Q} = \mathbf{J}(0)^\top\mathbf{J}(0) + \sum_{i=0}^{mn} \frac{\partial^2 e_i(\mathbf{x})}{\partial \mathbf{x}^2}|_{\mathbf{x}=0} e_i(0) \tag{4.11}$$

while the pose update results in :

$$\widehat{\mathbf{T}} \longleftarrow \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}) \tag{4.12}$$

Newton's method results in quadratic convergence of the cost function around $\mathbf{x} = 0$. Moreover, depending on the convexity of $\mathfrak{F}(\mathbf{x})$, the global minimum of the cost function can be found in minimal number of iterations (for *e.g.*, 2,3 iterations with iter frame displacements). For the case where $\mathfrak{F}(\mathbf{x})$ is non-convex, convergence problems occur if $\mathbf{Q}$ is not positive definite. On the other hand, Newton's method requires the expensive computation of the Hessian matrix. Several methods exist in literature to approximate matrix $\mathbf{Q}$ with a positive definite matrix $\widehat{\mathbf{Q}}$ which comes to the first order approximation of equation (4.7). These are listed as follows:

- Gradient descent:
$$\mathbf{Q} \approx \widehat{\mathbf{Q}} = \alpha\mathbf{I}, \alpha > 0 \tag{4.13}$$
- Gauss-Newton:
$$\mathbf{Q} \approx \widehat{\mathbf{Q}} = \mathbf{J}(0)^\top\mathbf{J}(0) \tag{4.14}$$
- Levenberg-Marquardt:
$$\mathbf{Q} \approx \widehat{\mathbf{Q}} = \mathbf{J}(0)^\top\mathbf{J}(0) + \alpha\mathbf{I}, \alpha > 0 \tag{4.15}$$

The above-mentioned methods, require an initial guess of the pose and it happens that an initial solution does not give an optimum of $\mathbf{x}$ depending on how far the initialization has

been made with respect to the reference. To be able to recover a pose close to the solution, equations (4.4) to (**??**) are evaluated iteratively until a tolerated threshold in $\|\mathbf{x}\| < \epsilon$ is reached.

Second order approximation of the cost function (4.41) are usually not applied due to the computational cost involved in evaluating the Hessian matrix. In [Baker & Matthews 2004], a plethora of algorithms are discussed; forward additive, forward compositional(FC), inverse compositional (IC). In particular, FC shows the equivalence of pose compositions possible when increasing DOFs is tackled, while IC is an improved formulation of FC which is inclined on reducing the computational burden of the tracking algorithm by keeping the Jacobian constant throughout the registration process such that the following approximation holds:

$$\mathbf{J}(0) \approx \widehat{\mathbf{J}}, \tag{4.16}$$

where, $\widehat{\mathbf{J}}$ is the Jacobian computed from the parameters of the reference frame only and remains constant thoughout as long as the reference frame holds. The latter is computed once and for all out of the optimization loop. This is possible by inverting the role of the input(current frame) and the template (reference frame) of equation (4.4).

4.2.1.2 Efficient Second Order Minimization (ESM)

Proposed by [Malis 2004] and [Benhimane & Malis 2004], the approach analyses the problem of the Hessian matrix computation by providing an enhanced approximation as applied to equation (4.11) without the need of its explicit treatment. This is achieved by a first order approximation of $\mathbf{M}(0, \mathbf{x})$ around $\mathbf{x} = 0$ whilst keeping the important property of positive definiteness on $\mathbf{Q}$. This extrapolation can be written as follows:

$$\mathbf{M}(0, \mathbf{x}) = \mathbf{J}(\mathbf{x}) - \mathbf{J}(0) + \mathbf{O}(\|\mathbf{x}\|^{\mathbf{2}}) \tag{4.17}$$

Plugging (4.17) into (4.7), the cost function now relativises to:

$$\mathbf{e}(\mathbf{x}) = \mathbf{e}(0) + \frac{1}{2}(\mathbf{J}(0) + \mathbf{J}(\mathbf{x}))\mathbf{x} + \mathbf{O}(\|\mathbf{x}\|^{\mathbf{3}}) \tag{4.18}$$

Second order approximation holds when the state variable $\mathbf{x} = \widetilde{\mathbf{x}}$, leading to:

$$\mathbf{e}(\widetilde{\mathbf{x}}) \approx \mathbf{e}(0) + \frac{1}{2}(\mathbf{J}(0) + \mathbf{J}(\widetilde{\mathbf{x}}))\widetilde{\mathbf{x}} \tag{4.19}$$

Denoting, $\mathbf{J}_{esm} = \mathbf{J}(0) + \mathbf{J}(\widetilde{\mathbf{x}})$, the cost function to be minimized is then re-written as:

$$\mathfrak{F}(\mathbf{x}) = \frac{1}{2}\|\mathbf{e}(\mathbf{0}) + \mathbf{J}_{esm}\mathbf{x}\|^2, \tag{4.20}$$

its first order derivative at the local minimum $\mathbf{x} = \widetilde{\mathbf{x}}$ is given as:

$$\nabla_{\mathbf{x}}\mathfrak{F}(\triangle\,\mathbf{x}) = \mathbf{J}_{esm}^{\top}(\mathbf{e}(0) + \mathbf{J}_{esm}\widetilde{\mathbf{x}}) = 0, \tag{4.21}$$

where the normal equation is given by:

$$\widetilde{\mathbf{x}} = -(\mathbf{J}_{esm}^{\top}\mathbf{J}_{esm})^{-1}\mathbf{J}_{esm}^{\top}\mathbf{e}(0), \tag{4.22}$$

and the pose update follows suit as in equation (4.12). The mechanism of ESM is more costly as compared to IC with the computation of the Jacobian the the current state $\mathbf{J}(0)$ at each iteration. But the benefits are multifold; approximately twice as less number of iterations to convergence than IC increase robustness to subsampling and noise and an increase basin of convergence as compared to first order estimation techniques.

4.2.1.3 Spherical Photometric cost function

Following the L-K formulation of equation (4.4), the latter is extended in order to adapt to the set of augmented spheres introduced in chapter (3). The error equation now embedding a projection model, following [Meilland *et al.* 2011a], is written as:

$$\mathbf{e}(\mathbf{x}) = \boldsymbol{\mathcal{I}}(w(\mathbf{T}(\tilde{\mathbf{x}}); \boldsymbol{\mathcal{Z}}; \mathbf{p})) - \boldsymbol{\mathcal{I}}^{*}(\mathbf{I}; \mathbf{p}^{*}), \tag{4.23}$$

Using chain rule, the error derivative is analytically computed as follows:

$$\frac{\partial \mathbf{e}(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial \boldsymbol{\mathcal{I}}(w(.))}{\partial w}\frac{\partial w}{\partial \mathbf{x}} \quad \Longrightarrow \frac{\partial \boldsymbol{\mathcal{I}}(w(.))}{\partial w}\left\{\frac{\partial w}{\partial \mathbf{T(x)}}\frac{\partial \mathbf{T(x)}}{\partial \mathbf{x}}\right\}, \tag{4.24}$$

Therefore, the total Jacobian $\mathbf{J}_{\mathbf{x}}$ is composed of three terms such that:

$$\mathbf{J}_{\mathbf{x}} = \mathbf{J}_{\boldsymbol{\mathcal{I}}}\mathbf{J}_{w}\mathbf{J}_{\mathbf{T}}, \tag{4.25}$$

corresponding to the Image, Projection and the Pose respectively. The composition $\mathbf{J_w}\mathbf{J_T}$ gives rise to the geometric part of Jacobian while $\mathbf{J}_{\boldsymbol{\mathcal{I}}}$ corresponds to the photometric part.

From equation 4.25, $\mathbf{J}_{\mathbf{T}}$, defined by Lie Algebra is given by:

$$\mathbf{J}_{\mathbf{T}} = \begin{array}{c} \\ \frac{\partial}{\partial v_x} \\ \frac{\partial}{\partial v_y} \\ \frac{\partial}{\partial v_z} \\ \frac{\partial}{\partial \omega_x} \\ \frac{\partial}{\partial \omega_y} \\ \frac{\partial}{\partial \omega_z} \end{array} \begin{array}{c} \begin{array}{cccccccccccc} 0 & -\omega_z & -\omega_y & v_x & \omega_z & 0 & -\omega_x & v_y & -\omega_y & \omega_x & 0 & v_z \end{array}^{\top} \\ \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

$\mathbf{J}_w$ can be further decomposed into $\mathbf{J}_\Pi$ and $\mathbf{J}_R$ where $\mathbf{J}_R$ is the the derivative with respect to a rigid point in space $(X, Y, Z) \in \mathbb{R}^3$ and $\mathbf{J}_\Pi$ depends on the projection model of the reference image.

Therefore, from $\mathbf{P}' = \mathbf{RP} + \mathbf{t}$, the derivative w.r.t the 12 elements of the transformation matrix is given by:

$$\text{from } \mathbf{P}' = \begin{bmatrix} r_{11}X + r_{12}Y + r_{13}Z + t_x \\ r_{21}X + r_{22}Y + r_{23}Z + t_y \\ r_{31}X + r_{32}Y + r_{33}Z + t_z \end{bmatrix} \qquad \Longrightarrow$$

$$\mathbf{J}_\mathbf{R} = \frac{\partial \mathbf{P}'}{\partial \mathbf{T}} = \begin{matrix} \frac{\partial \mathbf{P}'_1}{\partial \mathbf{T}} \\ \frac{\partial \mathbf{P}'_2}{\partial \mathbf{T}} \\ \frac{\partial \mathbf{P}'_3}{\partial \mathbf{T}} \end{matrix} \begin{matrix} \begin{matrix} r_{11} & r_{12} & r_{13} & t_x & r_{21} & r_{22} & r_{23} & t_y & r_{31} & r_{32} & r_{33} & t_z \end{matrix} \\ \begin{pmatrix} X & Y & Z & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & X & Y & Z & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & X & Y & Z & 1 \end{pmatrix} \end{matrix}$$

To complete the geometric part, for a spherical perspective projection, the mapping from cartesian to spherical coordinates is given by:

$$\begin{bmatrix} \theta \\ \phi \\ \rho \end{bmatrix} = \begin{bmatrix} \arctan(Z/X) \\ \arctan(Y/\sqrt{X^2 + Z^2}) \\ \sqrt{X^2 + Y^2 + Z^2} \end{bmatrix}, \tag{4.26}$$

where, $\|\rho\| = 1$ for a unit sphere. Hence,

$$\mathbf{J}_\Pi = \begin{matrix} \theta' \\ \phi' \\ \rho' \end{matrix} \begin{matrix} \begin{matrix} \frac{\partial}{\partial X} & \frac{\partial}{\partial Y} & \frac{\partial}{\partial Z} \end{matrix} \\ \begin{pmatrix} \frac{-Z}{\sqrt{X^2+Z^2}} & 0 & \frac{X}{\sqrt{X^2+Z^2}} \\ \frac{-XY}{\rho^2\sqrt{X^2+Z^2}} & \frac{\sqrt{X^2+Z^2}}{\rho^2} & \frac{-YZ}{\rho^2\sqrt{X^2+Z^2}} \\ 0 & 0 & 0 \end{pmatrix} \end{matrix} \tag{4.27}$$

Finally, the photometric Jacobian $\mathbf{J}_{\boldsymbol{\mathcal{I}}}$, related to a pixel $\mathrm{p} = (u, v)$ is given by:

$$\begin{aligned} \nabla_u \boldsymbol{\mathcal{I}}(u, v) &= \frac{\boldsymbol{\mathcal{I}}(u + \delta u, v) - \boldsymbol{\mathcal{I}}(u - \delta u, v)}{2\delta u} \\ \nabla_v \boldsymbol{\mathcal{I}}(u, v) &= \frac{\boldsymbol{\mathcal{I}}(u, v + \delta v) - \boldsymbol{\mathcal{I}}(u, v - \delta v)}{2\delta v} \\ \nabla_z \boldsymbol{\mathcal{I}}(u, v) &= 0 \end{aligned}$$

$$\therefore \qquad \mathbf{J}_{\boldsymbol{\mathcal{I}}} = \begin{bmatrix} \nabla_u \boldsymbol{\mathcal{I}}(u, v) \\ \nabla_v \boldsymbol{\mathcal{I}}(u, v) \\ 0 \end{bmatrix}^\top \tag{4.28}$$

After composition of the multiple Jacobians described above, the final Jacobian is the one described in equation (4.25). Consequently, the cost function is devised as follows :

$$\mathfrak{F}_I = \frac{1}{2}\sum_{i}^{k}\left\|\boldsymbol{\mathcal{I}}(w(\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x});\boldsymbol{\mathcal{P}}_i)) - \boldsymbol{\mathcal{I}}^*(w(\mathbf{I};\boldsymbol{\mathcal{P}}_i^*))\right\|^2, \tag{4.29}$$

where $w(.)$ is the warping function that projects a 3-D point $\boldsymbol{\mathcal{P}}_i$. Over here, $\boldsymbol{\mathcal{P}}_i$ encapsulates the 3-D projection and the pose $\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})$ is an approximation of the true transformation $\mathbf{T}(\tilde{\mathbf{x}})$.

4.2.2 Rigid Body Motion

To accurately recover the position of the RGBD spheres with respect to one-another, the pose $\mathbf{x}$ is parametrised using 6 DOFs – decomposed in two respective components; rotation and translation. Considering an RGBD sphere $\boldsymbol{\mathcal{S}} = \{\boldsymbol{\mathcal{I}}, \boldsymbol{\mathcal{D}}\}$ as defined in section (3.7.1), the objective is now to extract a transformation matrix between a reference sphere and the next one. The localisation problem is then similar to estimating the relative transformation $\widehat{\mathbf{T}}$ between the two consecutive spheres. The principle of rigid body motion is applied where the transformation of a point tethered to a coordinate frame represent the whole compact body motion. For any point pair lying on the body, metric properties such as distances and orientation are preserved. This kind of body motion, discussed subsequently forms part of the special euclidean group $\mathbb{SE}(3)$.

Inter-frame incremental displacement is further defined as an element of the Lie groups applied on the smooth differential manifold of $\mathbb{SE}(3)$ [Blanco 2010], also known as the group of direct affine isometries. Motion is parametrized as a twist (a velocity screw motion around an axis in space), denoted as $\mathbf{x} = \{[\boldsymbol{\omega}, \boldsymbol{\upsilon}] | \upsilon \in \mathbb{R}^3, \hat{\omega} \in so(3)\} \in \mathfrak{se}(3)$: $\boldsymbol{\omega} = \begin{bmatrix}\omega_x & \omega_y & \omega_z\end{bmatrix}$, $\boldsymbol{\upsilon} = \begin{bmatrix}\upsilon_x & \upsilon_y & \upsilon_z\end{bmatrix}$, with $so(3) = \{\hat{\omega} \in \mathbb{R}^{3\times3} | \hat{\omega} = -\hat{\omega}^\top\}$, where ω and υ are the angular and linear velocities respectively . The reconstruction of a group action $\widehat{\mathbf{T}} \in \mathbb{SE}(3)$ from the twist consists of applying the exponential map using Rodriguez formula [Ma *et al.* 2004]. Thereon, $\widehat{\mathbf{T}}$ is denoted as the transformation (pose) recovered between the current frame $\boldsymbol{\mathcal{I}} \in \mathbb{R}^{m\times n}$ observed at time t and the reference frame $\boldsymbol{\mathcal{I}}^* \in \mathbb{R}^{m\times n}$.

The output of equation (4.10) above gives rise to an instantaneous angular and translational velocities, corresponding to the camera motion $\mathbf{x} = [\boldsymbol{\omega}, \boldsymbol{\upsilon}]$ computed at each iteration of the cost function. In order to recover the instantaneous rotation and translation in cartesian space, $\mathbf{x}$, is integrated over time with an integration period of $\delta t = 1$:

$$\mathbf{x} = \int_0^1 (\boldsymbol{\omega}, \boldsymbol{\upsilon}) dt \in \mathfrak{se}(3) \tag{4.30}$$

The exponential map provides a way of performing the integral above so as to extract the transformation matrix. In literature, $\mathbf{x} \in \mathbb{R}^6$ is also known as a velocity screw or a twist ξ, when concatenated in a 4×4 matrix. Rotational rigid-body motion in $\mathbb{SE}(3)$ can be represented by a 3×3 matrix $\mathbf{R} \in SO(3)$. Given a trajectory $\mathbf{R}(t) : \mathbb{R} \rightarrow SO(3)$ that

describes a continuous rotational motion, the rotation must satisfy the following constraint:

$$\mathbf{R}(t)\mathbf{R}(t)^\top = \mathbf{I}$$

Computing the derivative of the above equation with respect to time t :

$$\dot{\mathbf{R}}(t)\mathbf{R}^T(t) + \mathbf{R}(t)\dot{\mathbf{R}}^\top(t) = 0 \implies \dot{\mathbf{R}}(t)\mathbf{R}^\top(t) = -(\dot{\mathbf{R}}^T(t)\mathbf{R}(t))^\top, \tag{4.31}$$

where $\dot{\mathbf{R}}(t)\mathbf{R}^T(t) \in \mathbb{R}^{3\times 3}$ is a skew symmetric matrix. From lemma [Ma *et al.* 2004], there exists a vector $\omega(t) \in \mathbb{R}$ such that

$$\dot{\mathbf{R}}(t)\mathbf{R}^\top(t) = \widehat{\omega}(t)$$

Multiplying both sides by $\mathbf{R}(t)$ and assuming that $\widehat{\omega}$ is constant in times yields:

$$\dot{\mathbf{R}}(t) = \widehat{\omega}\mathbf{R}(t). \tag{4.32}$$

Interpreting $\mathbf{R}(t)$ as the state transition matrix for the following linear ordinary differential equation (ODE):

$$\dot{x}(t) = \widehat{\omega}x(t), \qquad x(t) \in \mathbb{R}^3. \tag{4.33}$$

Then, the solution to the above first order ODE is given by:

$$x(t) = e^{\widehat{\omega}t}x(0), \tag{4.34}$$

where, $e^{\widehat{\omega}t}$ is the matrix exponential

$$e^{\widehat{\omega}t} = I + \widehat{\omega}t + \frac{(\widehat{\omega}t)^2}{2!} + \cdots + \frac{(\widehat{\omega}t)^n}{n!} + \cdots. \tag{4.35}$$

Assuming an initial condition of $\mathbf{R}(0) = I$,

$$\mathbf{R}(t) = e^{\widehat{\omega}t} \tag{4.36}$$

Coming back to our problem of pose estimation, $\mathbf{T}(\mathbf{x}) = [\mathbf{R} \quad \mathbf{t}] \in \mathbb{SE}(3)$, is recovered by applying the exponential map of the twist:

$$\mathbf{T}(\mathbf{x}) = e^{[\mathbf{x}]_\wedge} \tag{4.37}$$

where,

$$[\mathbf{x}]_\wedge = \begin{bmatrix} \widehat{\boldsymbol{\omega}} & \boldsymbol{\upsilon} \\ \mathbf{0} & 0 \end{bmatrix} \in \mathfrak{se}(3), \quad \widehat{\boldsymbol{\omega}} = \begin{bmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{bmatrix}, \boldsymbol{\upsilon} = \begin{bmatrix} \upsilon_x \\ \upsilon_y \\ \upsilon_z \end{bmatrix} \tag{4.38}$$

By using Rodrigues' formula coupled with additional properties of the matrix exponential,

the following relationship is established:

$$e^{[\mathbf{x}]_\wedge} = \begin{bmatrix} e^{\widehat{\omega}} & \frac{I - e^{\widehat{\omega}} v + \omega\omega^T v}{\|\omega\|} \\ 0 & 1 \end{bmatrix}, \quad \text{if} \quad \omega \neq 0, \tag{4.39}$$

from where, Rodrigues' formula for a rotation matrix, given $\omega \in \mathbb{R}^3$, the matrix exponential $\mathbf{R} = e^{\widehat{\omega}}$ is given by:

$$e^{\widehat{\omega}} = I + \frac{\widehat{\omega}}{\|\omega\|}\sin(\|\omega\|) + \frac{\widehat{\omega}^2}{\|\omega\|^2}(1 - \cos(\|\omega\|)). \tag{4.40}$$

4.2.3 Weighting Functions

Though direct methods are intrinsically robust with the amount of information redundancy present, a global minimum of the cost function at the solution is not always guaranteed. Iterative optimisation techniques are quite sporadic and may easily deviate form the solution when errors are pronounced. Error discrepancies occur under phenomena such as occlusions, dynamic foreground objects, non rigid entities such as vegetation for *e.g.*, or sensor noise are some aberrations which might occur between scenes. Therefore, to improve estimation and avoid local minima, robust penalty functions are rather implemented to penalise error functions. Their immediate effects downweight the contributions of high errors whilst favouring entities with low errors to drive optimisation. This method is also known as iterative reweighted least squares (IRLS).

In the presence of a suitable penalty function, the cost function is re-written as:

$$\underset{x \in \Re^n}{\operatorname{argmin}} \mathfrak{F}(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{mn} \Psi\big(e_i(x)\big)\big(e_i(x)\big)^2 \equiv \frac{1}{2}\|\mathbf{e}(\mathbf{x})\|_\Psi^2, \tag{4.41}$$

where, $\Psi\big(e_i(x)\big)$ is just a scale factor for the corresponding residual $e_i(x)$. The weights are then incorporated in the normal equation similar to (4.22):

$$\widetilde{\mathbf{x}} = -\big(\mathbf{J}^\top \mathbf{D} \mathbf{J}\big)^{-1} \mathbf{J}^\top \mathbf{D} \mathbf{e}, \tag{4.42}$$

where $\mathbf{D}$ is a diagonal matrix of size $mn \times mn$:

$$\mathbf{D} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix} \tag{4.43}$$

Coming back to [Lucas & Kanade 1981], a quadratic cost function is penalised assuming a Gaussian distribution over the error likelihood *i.e.* $p(e_i|\mathbf{x}) = \mathcal{N}\big(e_i(\mathbf{x}), 0, \sigma\big)$, where the influence function $\Psi(y) = \frac{y}{\sigma^2}$ and hence the weight is deduced as $w_i = \frac{1}{\sigma^2}$. Practically, the error does not follow a simple Gaussian approximation and therefore, σ has to be found

using more robust tools. A common weighting function devised by [Huber 1981] associates a confidence level $w_i \in [0;1]$ to each pixel $\mathbf{p}_i$ such that:

$$w_i = \frac{\Psi(\delta_i/\sigma)}{\delta_i/\sigma}, \quad \Psi(u) = \begin{cases} u, & \text{if} |u| \leq a \\ a\frac{u}{|u|}, & \text{if} |u| > a, \end{cases} \tag{4.44}$$

where, δ_i is centred around the residue by $\delta_i = e_i - \text{Median}(e)$. The value of a is set to 1.345 for 95% confidence level. σ us then robustly computed using the Median Absolute Deviation (MAD), defined as $\sigma = c(|\delta_i - \text{Median}(\delta)|)$, where $c = 1.4826$ for a normal distribution.

Considering a robust weighting function as outlined above, the photometric cost function introduced in section 4.2.1.3 now relates to:

$$\mathfrak{F}_I = \frac{1}{2}\sum_{i}^{k} \boldsymbol{\Psi}_{hub} \left\| \boldsymbol{\mathcal{I}}(w(\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}); \boldsymbol{\mathcal{P}}_i)) - \boldsymbol{\mathcal{I}}^*(w(\mathbf{I}; \boldsymbol{\mathcal{P}}_i^*)) \right\|^2, \tag{4.45}$$

where, $\boldsymbol{\Psi}_{hub}$ is a robust weighting function on the error given by Huber's M-estimator [Huber 1981]. The latter plays an important role in reducing the effect of outliers by measuring the similarity between two corresponding pixels. Hence the weight computed accommodates partially the uncertainty associated to each pairing between a reference and a current frame.

4.2.4 Information Selection

In order to estimate displacement between two frames, a set of correspondences between them has to be found to constrain the motion model $(\mathbf{R}, \mathbf{t})$ efficiently. This is a vital step in visual odometry as bad feature matches lead to pronounced deviation from the real motion. In literature, two mainstreams are identified: the first one based on feature extraction while the second one uses dense (correspondence-free) methods [Fraundorfer & Scaramuzza 2012].

Both approaches exhibit their advantages and inconveniences. The former, based on point feature detection needs to undergo an identification phase where primitives such as blobs, corners, lines or edges are usual candidates. Good features are characterized in terms of several properties such as stability, computational efficiency, distinctiveness or invariance to geometric and photometric changes.

On the other hand, dense methods make use of the entire photometric and geometric information content for tracking. Along that streamline, [Dellaert & Collins 1999] argued that instead of using all the information content which is computation intensive, selection of a subset of good pixels that yield enough information about the 6 degrees of freedom (DOF) state vector could considerably reduce computational cost as well as data redundancy without compromising on the accuracy of the estimation.

Since direct methods rely on optimisation techniques based on the gradient of the cost

function, textureless image regions do not contain enough information (e.g uniform wall surface) and localisation problems. In fact, if the photometric gradient $\nabla_{\mathbf{p}_i}\boldsymbol{\mathcal{I}}(\mathbf{p}_i) = 0$, the ith line of the jacobian matrix $\mathbf{J}_i$ will contain only zeroes and hence do not influence pose estimation. Therefore, ignoring these uninformative pixels during the warping phase definitely helps in terms of robustness as well as computation time (when the pseudo inverse of equation 4.42) is evaluated). Therefore, a naive approach used to decrease computation cost is to sort out the photometric gradient as applied in [Baker & Matthews 2001] as follows:

$$i = \underset{i}{\operatorname{argmax}} ||\nabla\boldsymbol{\mathcal{I}}(i)|| \tag{4.46}$$

However, a selection based on photometry only may favour certain DOFs at the expense of others leading to less precise motion estimation. In this context, [Meilland *et al.* 2010] proposed an improved selection algorithm which relied on finding the most informative pixel subset based on the decomposition of the Jacobian matrix obtained from equation (4.45). Since the algorithm forms the backbone of our odometry technique used for pose graph building, in this section, we shall subsequently give an elaborate description of the saliency map.

The Saliency Map The photometric jacobian defined by the Normal Flow Constraint (NFC) as in equation(4.25) can be decomposed into its six DOFs as follows:

$$\mathbf{J} = \begin{bmatrix} \mathbf{J_1} & \mathbf{J}_2 & \mathbf{J}_3 & \mathbf{J}_4 & \mathbf{J}_5 & \mathbf{J}_6 \end{bmatrix}, \tag{4.47}$$

where, each column of $\mathbf{J} \in \mathbb{R}^{mn\times 1}$ contains the gradient associated to each DOF and can be seen as a saliency map once the elements are rearranged in matrix form. Figure 4.2 shows six images obtained from a synthesized sphere. The brighter the pixel value, the bigger is its corresponding gradient value. Images of the first row $(\mathbf{J}^1, \mathbf{J}^2, \mathbf{J}^3)$ showing the decomposition correspond to the translational motion while the second row $(\mathbf{J}^4, \mathbf{J}^5, \mathbf{J}^6)$ illustrates the rotational motion.

The objective is to extract a subset $\overline{\mathbf{J}} = [\overline{\mathbf{J_1}}\ \overline{\mathbf{J_2}}\ \overline{\mathbf{J_3}}\ \overline{\mathbf{J_4}}\ \overline{\mathbf{J_5}}\ \overline{\mathbf{J_6}}] \subset \mathbf{J}$, of dimensions $\mathbf{p} \times 6$, with $\mathbf{p} \ll mn$, consists of pixels which best condition each DOF of matrix $\mathbf{J}$. The algorithm then solves iteratively the following function:

$$\overline{\mathbf{J}} = \underset{i}{\operatorname{argmin}} \big(|\mathbf{J}_i^j| \diagdown \widetilde{\mathbf{J}}\big), \tag{4.48}$$

which corresponds to selecting each line of the original matrix $\mathbf{J}$ relating to the best gradient of the j^{th} column (j^{th}DOF). $\widetilde{\mathbf{J}} \subset \overline{\mathbf{J}}$ consists of an intermediate subset of lines in $\mathbf{J}$ which have already been selected and $\diagdown\widetilde{\mathbf{J}}$ excludes the line which has already been picked out. We hereby outline the backbone of the algorithm which has been implemented in the system and is also illustrated in figure 4.3:

- Decomposition of the Jacobian $\mathbf{J}$ gives in itself a saliency map pertaining to each

DOF of $\mathbf{x}$

- Columnwise, $\mathbf{J}$ is sorted out for the best pixel which is indexed in decreasing order of magnitude
- Each DOF is then looked-up for the best ranked pixel i in ascending order and the i^{th} line is lifted to a new table as shown in figure 4.3
- In case a particular pixel gives the best indexing in more than one DOF and given that it has already been selected, we proceed to the second best pixel and so on.
- The selection process is performed iteratively until all the pixels have been successfully sorted out
- The final result is a table of best ranked pixels with respect to their response to a particular DOF

Therefore, instead of using all the pixels from the intensity and depth map, a wise selection of the top 10-20% of the pixels are used for registration. Making use of the partial RGB-D structure of the spheres through the saliency map leads us to term *Semi-Dense VO*.

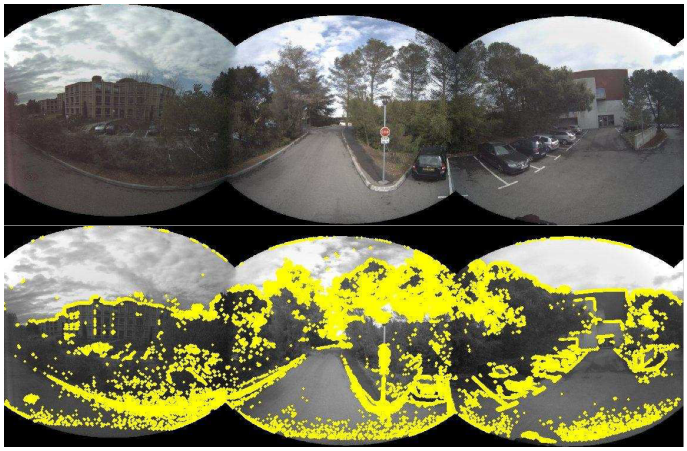

Figure 4.4: *Application of saliency map using top 5% of saliency map corresponding to around 68K pixels*

4.2.5 Multi-Pyramid resolution

One of the major inconvenience of direct optimisation techniques is that a good initialisation of the pose $\widehat{\mathbf{T}}$ which is close to the solution $\mathbf{T(\tilde{x})}$ is required so that the cost function closes down rapidly into the convergence domain and eventually to the solution. In order to improve the domain of convergence, a multi-resolution pyramidal approach is often employed. This consists of constructing a pyramid of N filtered and sub-sampled images by a factor of two [Burt & Adelson].

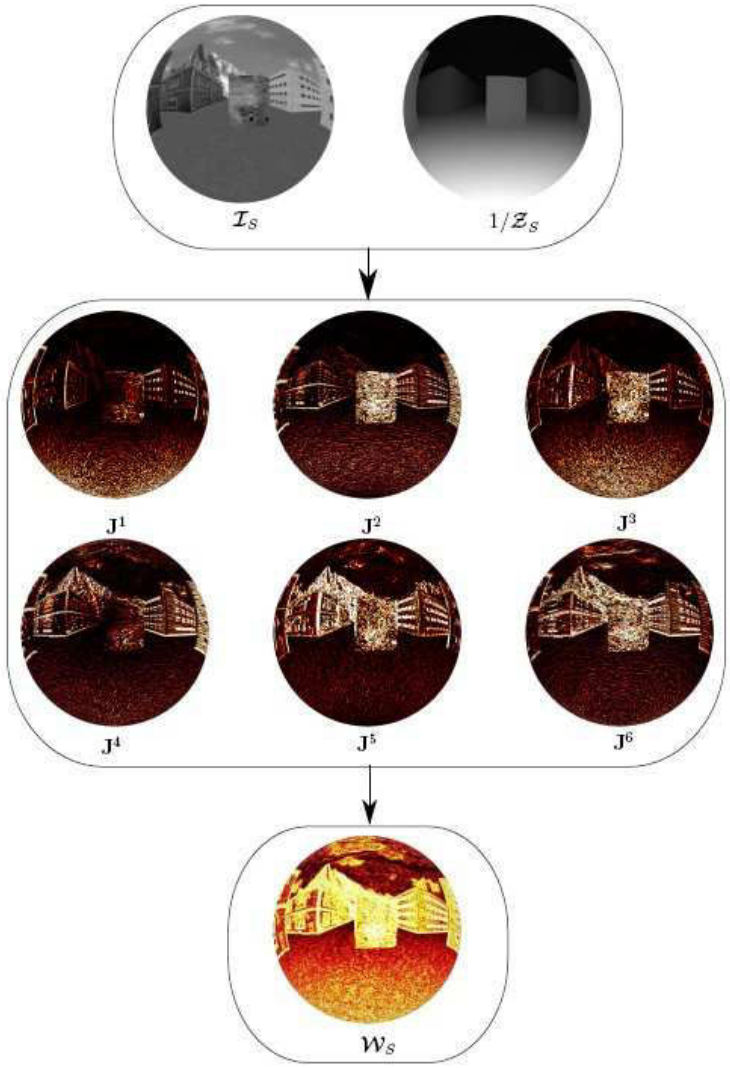

Figure 4.2: *Jacobian decomposition to saliency map*

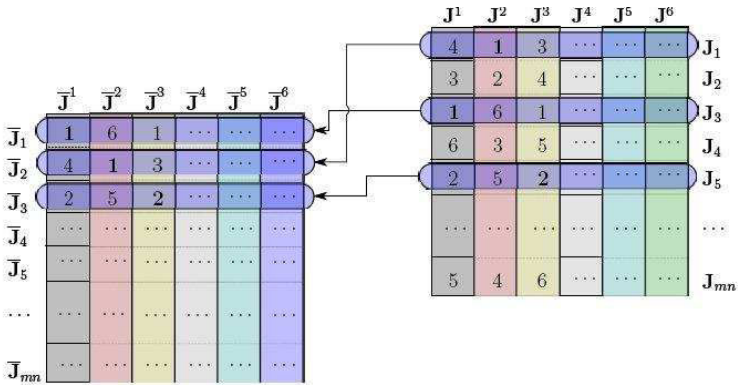

Figure 4.3: *Sorting out the saliency table*

Image level $k+1$ of the pyramid is obtained by successively subsampling the k^{th} image level $\boldsymbol{\mathcal{I}}^k$, convoluted by the following Gaussian kernel:

$$\boldsymbol{\mathcal{I}}^{K+1}(\mathbf{q}) = (\boldsymbol{\mathcal{I}}^k(\mathbf{p}) \otimes \mathbf{G})w(\mathbf{q}), \forall \mathbf{q} = 2(\mathbf{p}-1), \mathbf{p} \in \mathbb{N}^2, \tag{4.49}$$

where the function $w(\mathbf{q})$ selects pixel pair coordinates of the convoluted image $(\boldsymbol{\mathcal{I}}^k(\mathbf{p}) \otimes \mathbf{G})$. The Gaussian kernel $\mathbf{G}$ is defined by:

$$\mathbf{G} = \frac{1}{256}\begin{bmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 16 & 24 & 16 & 4 \\ 6 & 24 & 36 & 34 & 6 \\ 4 & 16 & 24 & 16 & 4 \\ 1 & 4 & 6 & 4 & 1 \end{bmatrix}$$

Level 0 of the pyramid correspond to the original image $\boldsymbol{\mathcal{I}}^0 \in \mathbb{R}^{m \times n}$. Successive pyramid levels are obtained from equation 4.49 up to level $N-1$, where the image $\boldsymbol{\mathcal{I}}^{N-1}$ is of dimensions $\frac{m}{2^{N-1}} \times \frac{n}{2^{N-1}}$. During image registration, the pyramids pertaining to the spherical photometric images $\boldsymbol{\mathcal{I}}$ and $\boldsymbol{\mathcal{I}}^*$ are first constructed.

As for the depth map $\boldsymbol{\mathcal{D}}$, subsampling by a factor of 2 is carried out without applying filtering in order to preserve its geometric content. However, it is worth noting that an alternative technique was recently applied in [Schöps *et al.* 2014] where their representation constituted of an inverse depth map $\boldsymbol{\mathcal{D}} : \Omega_{\boldsymbol{\mathcal{D}}} \rightarrow \mathbb{R}^+$ and its corresponding inverse variance map $V :\rightarrow \mathbb{R}^+$, where $\Omega_{\boldsymbol{\mathcal{D}}}$ contains all pixels with valid depth hypothesis. Consequently, the depth maps are down sampled using a weighted average of the inverse depth map as follows:

$$\boldsymbol{\mathcal{D}}_{l+1}(\mathbf{x}) = \frac{\sum_{\mathbf{x}' \in \Omega_\mathbf{x}} \frac{\boldsymbol{\mathcal{D}}_l(\mathbf{x}')}{V_l(\mathbf{x}')}}{\sum_{\mathbf{x}' \in \Omega_\mathbf{x}} \frac{1}{V_l(\mathbf{x}')}} \tag{4.50}$$

$$V_{l+1}(\mathbf{x}) = \frac{|\Omega_\mathbf{x}|}{\sum_{\mathbf{x}' \in \Omega_\mathbf{x}} \frac{1}{V_l(\mathbf{x}')}}, \tag{4.51}$$

where, $\Omega_\mathbf{x}$ denotes the set of valid pixel $\mathbf{x}$ at the next higher resolution. Averaging the inverse depth map using this technique better helps the photometric cost function, but not suitable for reconstruction purposes as it creates undesirable effects around depth discontinuities.

The multi-resolution minimisation algorithm begins at the $\mathrm{N}^{th}-1$ scale, corresponding to images of least dimensions with minimal content details. After convergence, the registration result is then injected to initialise the next pyramid level and optimisation is carried out again. This process repeats itself until the base level 0 is reached which corresponds to the largest resolution, hence the best content-wise precision. With this approach, a faster convergence of the cost function is achieved whilst avoiding local minima suppressed by the Gaussian filter. Furthermore, with this approach, larger interframe displacements are minimised at lower cost on the least resolution with better precision achieved on the biggest resolution, hence improving computation time. The figure 4.5 below illustrates the concept

based on three levels.

Figure 4.5: *Multi pyramid resolution*

4.3 Geometric constraint for motion estimation

So far, registration techniques elaborated in previous sections are categorized as 2D-3D correspondences where minimisation occurs in the image reprojection errors. Though this technique is more accurate, it does have certain limitations when exposed to lighting conditions. Cases where image based tracking fail are mostly due to unobservability in complete darkness; low level dynamic lighting scenarios exhibited in a living room for example, or cases of large bright spots attributed to TV/LCD screens, window panes exposed to bright sunlight. Moreover, scenes lacking extracted lines and curves are not easily handled with these approaches. Additionally, correction for large 3D pose errors happens to be more ambiguous due to significant pose errors [Zhao *et al.* 2005]. Geometric constraints, though less accurate than NFC [Scaramuzza & Fraundorfer 2011], is a good alternative to address the above mentioned limitations.

With the advent of consumer depth cameras such as the Microsoft Kinect or the Asus Xtion Pro structured light devices, depth information sensing has become more and more common. Theses classes of RGB-D cameras produce an active sensing range of 0.4 to 5 metres. However, beyond its upper limit, the measurements are not quite reliable. Different experimental set-ups seek to approximate static and dynamic errors so as to suppress the effect of sensor noise [Dryanovski *et al.* 2013][Khoshelham & Elberink 2012][Park *et al.* 2012]. Designed for the gaming industry, these sensors provide an important functionality in the acquisition of relatively high quality 3D range information in real time at low cost.

Consequently, with technological advancements, robotics research community have been given a big boost in using depth sensors which are now becoming an integral part of perception functionalities in intelligent mobile robots. Applications ranges from indoor wheeled robots [Endres *et al.* 2014] to flying robots [Kerl *et al.* 2013a] with the main objective including robot navigation and map building capabilities [Thrun *et al.* 2005]. In computer graphics community, scan alignment used for model reconstruction pipeline is an integral component in augmented reality applications for the entertainment industry as well

as the medical sector where the technology is implemented in computer assisted surgery. An interesting application related to digital archaeology can be found in [Levoy *et al.* 2000] whose aim was to conceive a 3D archive of museum sculptures and statues.

4.3.1 Direct Depth Map Alignment

In geometric motion estimation, features extracted from the 3D scene structure are used to compute a camera motion $\mathbf{T}(\tilde{\mathbf{x}})$ obtained by aligning two 3D feature sets [Scaramuzza & Fraundorfer 2011]. The cost underlying the geometric/3D point set registration technique is given by the direct minimization of the euclidean distance function as follows:

$$\mathfrak{F} = \underset{\mathbf{T}(\tilde{\mathbf{x}})}{\operatorname{argmin}} \sum_{i}^{k} \eta_i \|\boldsymbol{\mathcal{P}}_i - w(\mathbf{T}(\tilde{\mathbf{x}}); \boldsymbol{\mathcal{P}}_i^*)\|_2 \tag{4.52}$$

A robust closed form solution is presented in [Haralick *et al.* 1989], in a similar way to section (4.2.3) , where η comes from M-estimators whose role is to improve the performance and stability of the pose estimation by lessening the effect of outliers. $\mathbf{T}(\tilde{\mathbf{x}})$ is however not computed on the $\mathbb{SE}(\mathbf{3})$ manifold as presented earlier but approximated using rotational constraints with Lagrangian minimisation. Figure (4.6) depicts the warping of a 3D point $\mathbf{p}^*$ from a reference depth map $\boldsymbol{\mathcal{Z}}^*$ to a current frame using the projection equation (3.40) and the warping function encapsulated in equation (4.52).

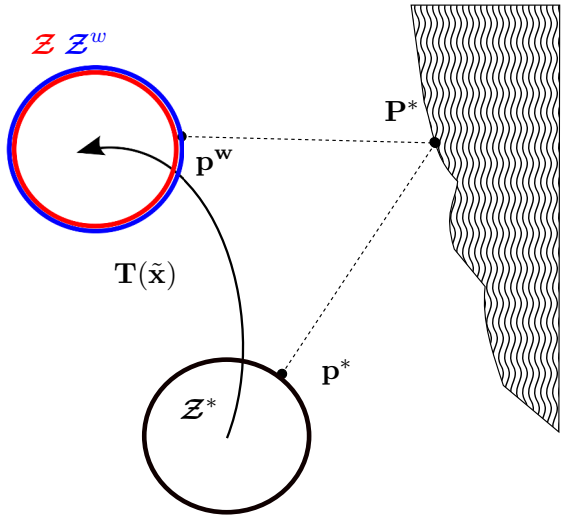

Figure 4.6: *3D geometric point projection and warping using depth maps*

The constraint presented above was first introduced by [Besl & McKay 1992] and [Chen & Medioni 1992]. The former (without weighting function) derived a generic formulation for data alignment which can be applied to geometric primitives such as point sets, line segment sets, curves and surfaces. The mechanism behind is famously known as Iterative Closest Point (ICP) which principally unfolds in two main stages:

1. finding correspondences between two datasets based on a proximity measure
2. optimisation over the parametrised motion parameters by applying a suitable cost function (*e.g.* using euclidean distance norm)

The above stages can further be exploded into the following core steps as presented in [Rusinkiewicz & Levoy 2001]:

1)Points Sampling: Instead of using all the available points [Besl & McKay 1992] of both point sets, (source and destination), sampling of point sets is rather desirable to extract points which are potentially matchable. Subsampling can be done in either the source set or both source and destination sets. Different techniques exist such as uniform or random sampling or selection of points with high intensity gradient for *e.g.*, the saliency algorithm presented earlier in section 4.2.4 therefore suits the purpose of points selection. [Rusinkiewicz & Levoy 2001] introduced a new sampling technique by choosing points with large distribution of normals among the choosen points so as to better constraint the rigid motion parameters. They further argued the impact of sampling strategy on performance, stability, computational burden and robustness of the cost function.

2) Matching samples: This step aims at finding point correspondences in a sample set. Unlike feature extraction and matching pipeline, ICP uses a simple euclidean distance heuristic to establish correspondences. However, closest point computation is computationally exhaustive and hence data structures such as a k-d tree [Zhang 1994]. Other data structures such as octrees [Steinbrücker *et al.* 2013] are also considered whilst [Newcombe *et al.* 2011] explored the use of volumetric signed distance function as an alternative to building accelerated data structures but operating in a dense reconstruction and tracking setting. In our work, the approach outlined in [Blais & Levine 1995] and [Neugebauer 1997] is rather preferred because it exploits the projective depth map structure which suits well our purpose since the latter is encoded on a spherical grid-like structure. Using the underlying structure, a point from the source depth map is projected onto the destination depth map by an estimated transform as shown in figure 4.6. The closest point is then obtained using a simple nearest neighbour search on the grid which is then taken to be the corresponding point. For small inter-frame displacements, this method holds well in practice for smooth surfaces although it is clearly invalid for regions with depth discontinuities.

3) Correspondence weighting and outlier rejection: To tackle the problem of data mismatching relating to scenarios of boundary points, depth discontinuities or noise, outliers are downweighted to limit their influence on the cost function. Over here, a multitude of techniques can be applied by defining a plethora of weighting functions [Haralick *et al.* 1989], similar to the ones described in section 4.2.3. Moreover, putative matches are rejected based on a predefined point-to-point metric distance threshold. The rejection phase is vital as it eliminates boundary points which systematically bias the estimated transform.

4) Error metric optimisation: In the last step of ICP, a parametrised motion transform between the point sets is then estimated by minimising over a suitable similarity metric. The original ICP algorithm developed by [Besl & McKay 1992] used a quadratic cost function with euclidean distance metric. [Zhang 1994] presented a robustified cost function in the same way as defined in equation 4.52 whilst [Chen & Medioni 1992] came up with a point to plane distance metric. This technique is given an in-depth treatment in our work and shall be discussed in the subsequent section. The latter forms the core of a dual photometric and geometric cost function investigated subsequently in section 4.4.

Recently, [Segal *et al.* 2009] proposed a Generalized ICP which is based on a Maximum Likelihood Estimation (MLE) probabilistic model. The derived model is generic in the sense that it can accomodate a point to point, a point to plane or a plane to plane model by tweaking the assumptions made on the covariance matrices of the point sets. All distance metrics can be solved using robust iterative non-linear minimisation techniques as discussed in section 4.2.1. [Lui *et al.* 2012] investigated an ICP flavour based on inverse depth parametrisation with bare and weighted point to point and point to plane variations and showed that an overall best performance is terms of speed and accuracy trade-off is manifested by the weighted point to plane error metric while the unweighted error metrics tend to converge slower with higher errors. Moreover, they also become unstable with larger interframe rotation discrepancies. A version of fast ICP was exposed in [Rusinkiewicz *et al.* 2002] by using the best permutations of the variants described in steps 1 to 4. In particular, their fast ICP pipeline, consisted of **i)** extraction of a subset of points from one mesh by random sampling **ii)** projective data association of the source points **iii)** an outlier rejection criterion based on a point to point distance threshold and finally **iv)** a point to plane error metric minimised over the euclidean norm. Their metholodogy has shown to be suitable for a high speed small baseline alignment of two projectively acquired depth map measurements.

On a different note, as pointed out earlier, the depth map representation that is used is encoded on a spherical grid with uniform sampling as introduced in section 3.7.2. From this representation, a point cloud structure is readily available. Overhere, we discuss, the advantages [Zhao *et al.* 2005] which comes part and parcel of the underlying structure. Notably, the use of 3D point clouds offers maximal flexibility meaning that this brute representation is not reliable over the extraction of geometric primitives or features. Therefore a prior preprocessing of line extraction or plane segmentation for example is avoided prior to registration. Furthermore, the advantages of 3D-3D registration are multifold; ability to handle large pose variation. This is vital since in practice, an initial estimate of the pose is only available and 2D representations could appear substantially different under large pose variations. Hence, alignment of 3D sensor data on a model is possible for unstructured areas such as vegetation. Finally, the complete geometry of the model environment can be dynamically built and improved with new incoming sensor data [Newcombe *et al.* 2011].

The flexibilities described above are what make ICP as the main engine for 3D data registration. ICP works best when complete 3D information of the environment is available

and convergence to a global minimum is achievable with a good initial estimate, together with a large superiority ratio of inliers to outliers. One of the reason for ICP failures are attributed to sparse representations where the number of inliers are highly diminished.

4.3.2 Point to plane registration

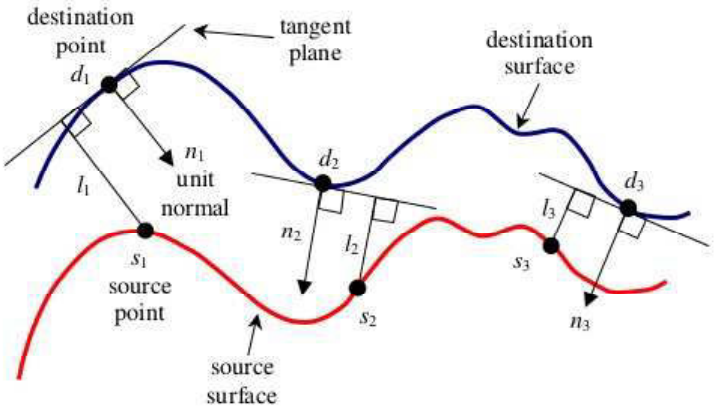

Figure 4.7: *Principle of ICP registration between two surfaces, courtesy of [Low 2004]*

Point-to-plane ICP, though slower than point-to-point, offers better convergence rates than the latter [Rusinkiewicz & Levoy 2001]. Moreover, point-to-point distance metric gets increasingly inaccurate when the orthogonality of viewing angle between the camera and the surface decreases and this is where uncertainties related to the measurement increases. Figure 4.7 above illustrates the working principle of point-to-point ICP. Starting with a raw depth map taken as the reference, an initial pose estimate $\mathbf{T}(\tilde{\mathbf{x}})$, the reference depth map is projected onto the current depth map. Combining projective data association with point-to-plane metric, the direct iterative alignment scheme leads us to the following error function:

$$\mathbf{e}(\mathbf{x}) = n_i^\top (\boldsymbol{\mathcal{P}}_i - \mathbf{T}(\tilde{\mathbf{x}})\boldsymbol{\mathcal{P}}_i^*), \tag{4.53}$$

with the jacobian computed as follows:

$$\frac{\partial \mathbf{e}(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial e(.)}{\partial \mathbf{T}(\tilde{\mathbf{x}})} \frac{\partial \mathbf{T}(\tilde{\mathbf{x}})}{\partial \mathbf{x}} \quad \Longrightarrow - \begin{bmatrix} n_X \\ n_Y \\ n_Z \end{bmatrix}^\top \begin{bmatrix} 1 & 0 & 0 & 0 & Z & -Y \\ 0 & 1 & 0 & -Z & 0 & X \\ 0 & 0 & 1 & Y & -X & 0 \end{bmatrix} \tag{4.54}$$

Since our RGB-D data is encoded on a spherical grid, the normal map n_k is computed from the two neighbouring vertices of the grid as follows, in a similar fashion to [Newcombe *et al.* 2011] as follows:

$$n_k(u,v) = \frac{\big(v_k(u+1,v) - v_k(u,v)\big) \times \big(v_k(u,v+1) - v_k(u,v)\big)}{\big\| \big(v_k(u+1,v) - v_k(u,v)\big) \times \big(v_k(u,v+1) - v_k(u,v)\big) \big\|_2} \tag{4.55}$$

Hence the objective function, is optimised using the Gauss-Newton descent approach as follows:

$$\mathfrak{F}_{icp} = \frac{1}{2}\sum_{i}^{k} \boldsymbol{\eta} \left\| n_i^{\top} \left(\boldsymbol{\mathcal{P}}_i - \mathbf{T}(\tilde{\mathbf{x}}) \boldsymbol{\mathcal{P}}_i^{*} \right) \right\|^2 \tag{4.56}$$

4.4 Motion Tracking englobing *Photo* + *Geo* constraints

After the well established theory of optical flow for motion estimation using Intensity based cost functions, the community has now turned to the fusion of both intensity and depth information for registration and tracking. This trend is recurrent especially due to the advent of consumer based RGB-D sensors such as Microsoft's Kinect or Asus's Xtion Pro. A breakthrough of this technique was devised by [Harville *et al.* 1999], who were among the first to formalize registration as a Brightness Change Constraint Equation (BCCE) and a Depth Change Constraint Equation (DCCE). They argued that tracking is best achieved with intelligent fusion of the two constraints mentioned above in order to reduce the effect of drift, occlusions or illumination changes.

Their work was further extended in [Rahimi *et al.* 2001], where a Maximum Likelihood (ML) based differential tracker was developed and the problem of drift and loop closure were also addressed. To improve the tracking performance, the measurement model incorporated the fusion of multiple frames. A similar formulation was proposed in [Wang *et al.* 2006], but encapsulated in a Bayesian framework. A 2D-3D pose estimation along with an intensity cost function helped to improve feature correspondence as well as drift reduction.

Recent works of the same domain includes that of [Newcombe *et al.* 2011] whereby a preliminary frame to frame registration is fed to a surface reconstruction module which improves the perceived model over time together with the estimated pose. The RGB-D slam framework of [Henry *et al.* 2012] used a variant of the ICP together with photometry. Their work also included a surfel representation of the environment to have a more compact representation of the information available. Other related works merging photometric and geometric information for Visual Odometry (VO) can be found in [Tykkälä *et al.* 2011], [Tykkälä *et al.* 2013], [Kerl *et al.* 2013a], [Meilland & Comport 2013a].

4.4.1 Cost Function Formulation

With the aim of robustifying the above cost function, a geometric point to plane constraint [Chen & Medioni 1992] is added to the equation (4.45), where the system is solved in a unified framework as follows:

$$\mathfrak{F}_S = \frac{\beta^2}{2}\|e_{\mathcal{I}}\|_{\Psi}^2 + \frac{\vartheta^2}{2}\|e_{\rho}\|_{\eta}^2, \tag{4.57}$$

which can be written in its explicit form:

$$\begin{aligned}\mathfrak{F}_S = & \frac{\beta^2}{2}\sum_i^k \boldsymbol{\Psi}_{HUB}\left\|\boldsymbol{\mathcal{I}}(w(\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x});\boldsymbol{\mathcal{P}}_i^*)) - \boldsymbol{\mathcal{I}}^*(w(\mathbf{I};\boldsymbol{\mathcal{P}}_i^*))\right\|^2 \\ + & \frac{\vartheta^2}{2}\sum_i^k \boldsymbol{\eta}_{HUB}\left\|n_i^\top(\boldsymbol{\mathcal{P}}_i - \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})\boldsymbol{\mathcal{P}}_i^*)\right\|^2,\end{aligned} \tag{4.58}$$

such that $\boldsymbol{\mathcal{P}} \in (X, Y, Z) \longrightarrow (\theta, \phi, \rho)$ and β, ϑ are tuning parameters to effectively balance the two cost functions. n_i^T is the normal map computed from the cross product of adjacent points on the grid structured depth map.

Since the unknown $\mathbf{x}$ is common in both parts of equation (4.58), the error function is stacked in a single vector computed simultaneously as shown:

$$\mathbf{e(x)}_S = \begin{bmatrix} \beta\boldsymbol{\Psi}_{HUB}\Big(\boldsymbol{\mathcal{I}}(w(\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x});\boldsymbol{\mathcal{P}}^*)) - \boldsymbol{\mathcal{I}}^*(\boldsymbol{\mathcal{P}}^*)\Big) \\ \vartheta\boldsymbol{\eta}_{HUB}\Big(n^\top(\boldsymbol{\mathcal{P}} - \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})\boldsymbol{\mathcal{P}}^*)\Big) \end{bmatrix} \tag{4.59}$$

The Jacobian matrix $\mathbf{J}_S$ is the total Jacobian relative to the augmented cost function defined above and is given as:

$$\mathbf{J}_S = \begin{bmatrix} \beta\mathbf{J}_{\mathcal{I}^*}\mathbf{J}_w\mathbf{J}_T \\ \vartheta n^T\mathbf{J}_D \end{bmatrix}, \tag{4.60}$$

Where, respectively, $J_{\mathcal{I}^*}$ is the jacobian w.r.t. the intensity, and $\mathbf{J}_w$ is the jacobian w.r.t. the warping function, $\mathbf{J}_T$ is the jacobian w.r.t. the pose and $\mathbf{J}_D$ is the jacobian w.r.t. the depth.

Similarly, the weighting function for each part of cost function is stacked in a block diagonal matrix where $\mathbf{D}_{\boldsymbol{\mathcal{I}}}, \mathbf{D}_{\boldsymbol{\mathcal{D}}} \in \mathbb{R}^{mn\times mn}$ are the confidence level in illumination and depth respectively for each corresponding feature pair:

$$\mathbf{D}_S = \begin{bmatrix} \mathbf{D}_{\boldsymbol{\mathcal{I}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{\boldsymbol{\mathcal{D}}} \end{bmatrix} \tag{4.61}$$

Linearization of the above cost function leads to a classic closed form solution given by an Iterative Least Mean Squares (ILMS) and the incremental motion $\mathbf{x}$ is given by the following expression:

$$\mathbf{x} = -(\mathbf{J}_S^T\mathbf{D}_S\mathbf{J}_S)^{-1}\mathbf{J}_S^T\mathbf{D}_S\mathbf{e}(\mathbf{x})_S \tag{4.62}$$

Using an iterative optimization scheme, the estimate is updated at each step by an homogeneous transformation:

$$\widehat{\mathbf{T}} \longleftarrow \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}), \tag{4.63}$$

where $\widehat{\mathbf{T}} = [\mathbf{R} \quad \mathbf{t}]$ is the current pose estimate with respect to the reference available from the previous iteration.

4.5 Keyframe-based Representation

When exploring vast scale environments, many frames sharing redundant information clutter the memory space considerably. The idea to select keyframes based on a predefined criteria happens to be very useful in the conception of a sparse skeletal pose graph. Furthermore, performing frame to frame registration introduces drift in the trajectory due to uncertainty in the estimated pose as pointed out in [Kerl *et al.* 2013a].

Therefore, in order to overcome this issue, frame to keyframe odometry is rather desirable. Common techniques applied constitute of introducing keyframes when two such frames share very few features between them as defined by the view clustering criteria of [Konolige & Bowman 2009], or a threshold on the number of features shared between a central frame and its corresponding adjacent frames [Royer *et al.* 2007]. [Strasdat *et al.* 2010], introduce a new frame whenever a certain distance threshold between camera poses is exceeded. [Wang *et al.* 2006] modeled a temporal criteria to take into account the interpose frame difference as well as feature overlap among them. On the other hand [Meilland *et al.* 2011a] used a selection criteria based on the Median Absolute Deviation (MAD) in intensity error between a reference and a current frame to reinitialize on a predefined threshold.

Recently, information theory [Kretzschmar *et al.* 2010] was introduced to prune out nodes with a minimal expected information gain. On a similar note, [Kim & Eustice 2013] set up a salient keyframe selection criteria based on the ratio between the covariance of the measurement and that of the innovation to encode the entropy between two corresponding nodes. However, this criteria is modeled based on the probabilistic framework of iSAM where the covariances are easily extracted. On the other hand, the criteria based on a differential entropy approach introduced by [Kerl *et al.* 2013a] was found to be more suitable for our system of geo-referenced spheres which will be discussed in section 4.5.2. Next, two criteria are elaborated in MAD and Entropy and the pros and cons of each one of them are underlined as they form an important aspect in useful keyframe selection and hence contribute a significant part in our pose graph representation.

Figure 4.8 illustrates our keyframe based representation that we deal with throughout the course of this work. It consists of a graph of nodes joined with edges established from VO. An agent A in the graph is able to localise itself with respect to the nearest keyframe it perceives using the information stored at that particular node.

4.5.1 Median Absolute Deviation

Perhaps the simplest technique of analysing derpersion between two images is to compute the statistical correlation between these two sets of values, or, analyse the residual error distribution between a reference sphere $\boldsymbol{\mathcal{S}}^*$ and that of a warped one, say $\boldsymbol{\mathcal{S}}^w$ represented in the same common frame. While the error's standard deviation gives an indication of the nodes' disparity, it is not robust and is greatly affected by outliers. On the other hand, the

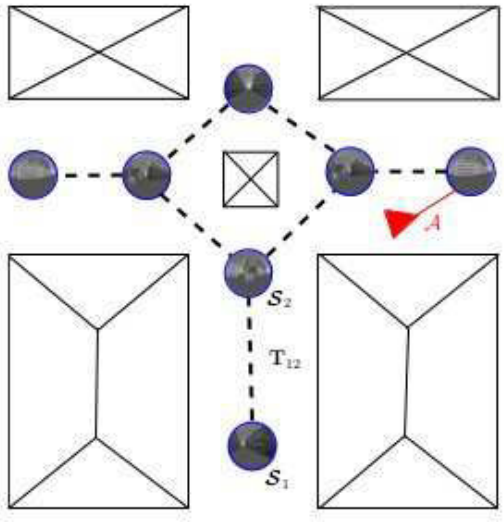

Figure 4.8: *Our keyframe-based representation of the environment, courtesy of [Meilland 2012]*

MAD offers better robustness to outliers with the criteria still holding good with at most 50% of outliers' presence [Twinanda *et al.* 2013]. The following equation defines the MAD as:

$$\sigma_{MAD} : median(|\mathbf{De}(\hat{\mathbf{x}}) - median(\mathbf{De}(\hat{\mathbf{x}}))|) > \lambda, \tag{4.64}$$

where $\mathbf{D}$ is a weighting matrix and $\hat{\mathbf{x}}$ is the pose estimate. The MAD is an increasing function capped at λ indicating the amount by which the photometric information of the scene has changed between the two spheres $\boldsymbol{\mathcal{S}}^*$ and $\boldsymbol{\mathcal{S}}^w$. Information changes are directly attributed to viewpoint changes inducing occlusions phenomena in the scene of the dynamic nature of the scene itself (for *e.g.* moving cars, objects, individuals ect ...). One major drawback of MAD is that it is univariate and can therefore be applied to only one entity. In the case of [Meilland *et al.* 2011a], it was applied to the intensity cost function. Consequently, the latter is highly affected by illumination changes in the scene and doesn't mean that the geometry of the scene has changed enough in order to proceed to a keyframe re-initialisation stage. Moreover, it is very much content based and the threshold has to be empirically set depending on the scene type.

4.5.2 Differential Entropy

With the aim of countering the shortcomings of the MAD, a new criteria is then developed in this section using the concept of entropy, following it's introduction in [Kerl *et al.* 2013b].

Differential entropy of a random variable $\mathbf{x}$ with dimensions n such that $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ is

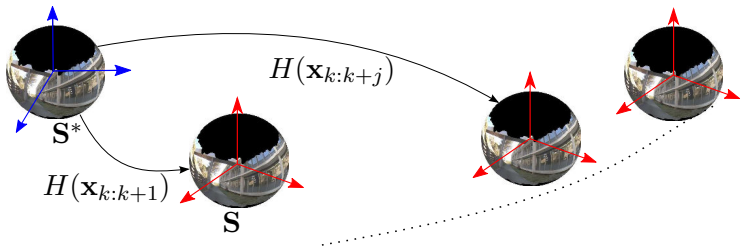

Figure 4.9: *Illustration of entropy ratio* α

defined as:

$$H(\mathbf{x}) = \frac{n}{2}(1 + \ln(2\pi)) + \frac{1}{2}\ln(|\Sigma|), \tag{4.65}$$

where, Σ is the covariance matrix of the estimate $\mathbf{x}$ which is obtained by the inverse of the Fisher Information matrix computed from the normal equations (4.62):

$$\Sigma = \left(\mathbf{J}_S^T \mathbf{D}_S \mathbf{J}_S\right)^{-1}, \tag{4.66}$$

which can be also decomposed into its components as follows:

$$\Sigma = \begin{bmatrix} \Sigma_\omega & \Sigma_{\omega,\nu}^\top \\ \Sigma_{\omega,\nu} & \Sigma_\nu \end{bmatrix} \tag{4.67}$$

The entropy ratio between a motion estimate $x_{k:k+j}$ from a reference frame k to a current frame $k+j$ is obtained by the following deduction:

$$\alpha = \frac{H(\mathbf{x}_{k:k+j})}{H(\mathbf{x}_{k:k+1})}, \tag{4.68}$$

where the denominator is just the entropy relative to the consecutive of the k^{th} frame in question. The greater the gap between the reference and the current frame, the greater is the pose uncertainty and the smaller is the value of α. Hence a preset on the value of α is used to decide whenever a keyframe needs to be inserted or not. Finally, α can be viewed as an abstraction of the pose's uncertainty encoded as a numerical value. Moreover, it does not depend on the illumination aspect of the sequence as the case of the MAD but depends on the quality of the geometry and hence the pose estimation step. Figure 4.9 illustrates how the criteria is applied on our database of augmented spheres.

4.6 Evaluation Metrics

The resulting map obtained from of a Visual SLAM system comes along with a generated trajectory. In order to validate the trajectory and thus the quality of the map obtained in some sense, the latter is generally compared to a ground truth map (hence the associated

trajectory). Accurate ground truth sequences are practically hard to obtain but if available, they provide a good evaluation.

Given a sequence of poses from an estimated trajectory $\mathrm{M}_1, \cdots\cdots \mathrm{M}_n \in \mathbb{SE}(3)$ and that of the ground truth defined as $\mathrm{N}_1, \cdots\cdots \mathrm{N}_n \in \mathbb{SE}(3)$. It is further assumed that the sequences are synchronised, having the same length with the same number of samples. However, in practice, this is not always true as the two different sequences may have dissimilar sampling rates, of different lengths or even missing data which would require and additional interpolation step as recalled in [Sturm *et al.* 2012]. Two common evaluation metrics mentioned in literature are the *relative pose error* (RPE) and the *absolute trajectory error* (ATE). RPE, which measures the local accuracy of the trajectory over a fixed time interval δ is obtained as follows:

$$\mathbf{E}_i^{\text{rpe}} = \left(\mathrm{N}_i^{-1}\mathrm{N}_{i+\delta}\right)^{-1}\left(\mathrm{M}_i^{-1}\mathrm{M}_{i+\delta}\right), \tag{4.69}$$

where i is the current time step. From the error vector computed above, its root means squared error (RMSE) over a sequence of n camera poses can be obtained as follows:

$$\mathbf{E}_{1:n}^{\text{rms}} = \sqrt{\frac{1}{n-\delta}\sum_{i=1}^{n-\delta} ||\mathbf{e}^\top \mathbf{E}_i^{\text{rpe}}||^2}, \tag{4.70}$$

where $\mathbf{e}^\top = [0 \;\; 0 \;\; 0 \;\; 1]$ is a row vector to extract the translational components of $\mathbf{E}$. Apart from the RMSE, other evaluations such as the mean error or the median error may also be applied which reduce the influence of ouliers. Rotational error, too can be evaluated but since the latter is strongly coupled to translation, it eventually appear in the translational error.

On the other hand, the *absolute trajectory error* (ATE) measures the global consistency of the estimated trajectory. This is obtained by comparing the absolute distances between the estimated and the ground truth trajectory. ATE is computed as follows:

$$\mathbf{E}_i^{\text{ate}} = \mathrm{N}_i^{-1}\mathbf{T}\mathrm{M}_i, \tag{4.71}$$

where $\mathbf{T}$ is the transformation which maps N onto M when both trajectories are not in the same reference frame. Similarly, the root mean square error is evaluated as follows:

$$\mathbf{E}_{1:n}^{\text{rms}} = \sqrt{\frac{1}{n}\sum_{i}^{n} ||\mathbf{e}^\top \mathbf{E}_i^{\text{ate}}||^2} \tag{4.72}$$

RPE considers explicitly both translational and rotational errors (formulation 4.69), while ATE considers only translational errors. Consequently, RPE is slightly greater than ATE or equal to for the case where rotational error is negligible. However, as mentioned earlier, both components – translation and rotation are highly correlated and hence discrepancies in rotation affect translation as well and hence captured by ATE. Ultimately, there is

no substantial difference in both metrics as pointed out in [Sturm *et al.* 2012]. In the following results section, ATE shall be used as the chosen criteria to compute the discrepancies between the ground truth and the VO generated trajectory.

4.7 Results and Discussion

4.7.1 Synthetic dataset:

Our first fold of experiments have been performed on two synthetic datasets; one with a sequence modelled with spherical illumination (Sph. Illum)and the other with diffuse illumination (Diff. Illum). Each set consists of spherical intensity and depth maps of size 640×480 generated from Sponza Atrium model [Meilland & Comport 2013b] and is provided with ground truth poses. The sequence is made of extended corridors, alleys, textured inner and outer buildings' surfaces as depicted in figure 4.10. Our algorithms have been thoroughly tested on this dataset of around 600 images in order to validate the convergence of the various cost functions as well as the keyframe criteria discussed in this chapter before moving on to real data. The permuted set of experimentations is defined as follows:

- Expt (a): Diff. illum + ESM + MAD
- Expt (b): Sph. + ESM +MAD
- Expt (c): Sph. ESM +ICP + MAD
- Expt (d): Sph. ESM +ICP + Entropy

Experiment	ATE/m	Nos. of Keyframes
Expt (a)	0.2480	23
Expt (b)	0.9501	109
Expt (c)	0.8189	93
Expt (d)	**0.2993**	**58**

Table 4.1: *Methods comparison*

Figures 4.11(a) to (d) illustrates the trajectories obtained from the different experimental sets mentioned above. Figure 4.11(a) refers to the trajectory of Expt(a) which is the ideal case of photometry and geometry. As illustrated, the trajectory obtained perfectly follows the ground truth trajectory with a sparse number of registered keyframes. On the other hand, figure 4.11(b) which relates to Expt(b) shows how the problem of illumination can heavily affect the trajectory. It is observed that along the trajectory, the NFC estimation function has not converged properly even though the maximum number of iterations were capped at 200 resulting in an accumulated drift in the direction of motion.

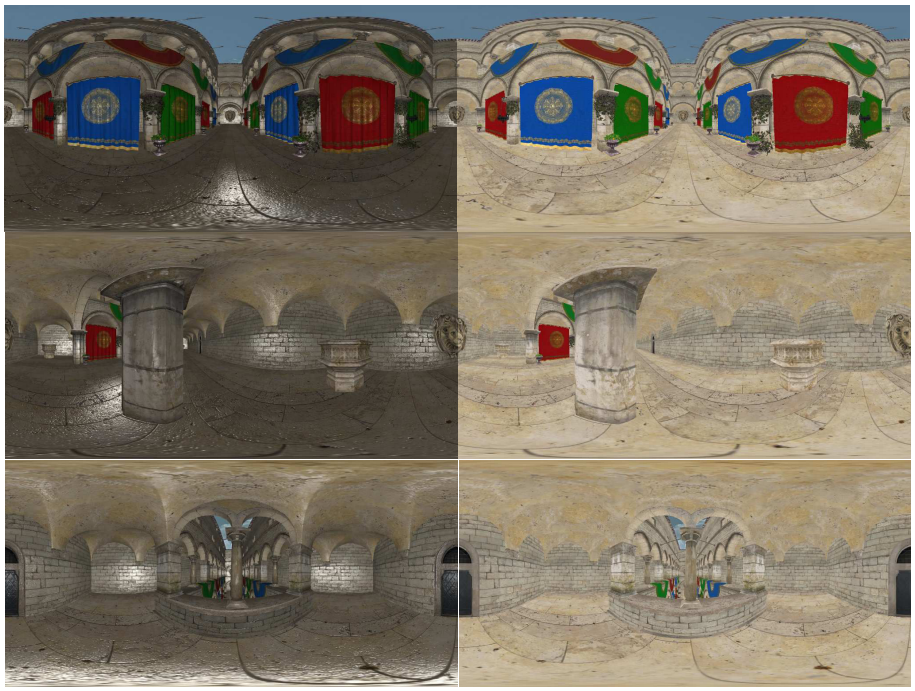

Figure 4.10: *Presentation of synthetic dataset with spherical (left column) and diffuse illumination (right column)*

The fact that no illumination model was considered in the NFC error function might suggest that the cost function runs into local minima and hence the discrepancy. The number of keyframes registered were 109 along a total trajectory of around 55m. Figure 4.11(c) refers to Expt(c) where the dual photometric and geometric cost function is tested with the MAD as Keyframe criteria. The trajectory obtained comparatively follows that of the ground truth but with accumulated drift which is inherently a VO problem. A plausible explanation would be that since the MAD is used as keyframe criteria, the graph is denser and error accumulation along the trajectory gets bigger which is then propagated across the whole chain. This hypothesis is finally confirmed in figure 4.11(d), whereby using less keyframes, the overall tracking drift is greatly reduced and the generated trajectory closes on that of the ground truth. The MAD is set to a heuristic value of 5, while the differential entropy criteria α is thresholded at 0.96. These two criteria cannot be directly compared as they are extracted from two very different entities. The MAD which varies on the illumination aspect of the scene is rather quickly reached. On the other hand, the entropy criteria varies rather on the uncertainty of the pose performs better than the former. Moreover, it does not require pre-tuning on a specific dataset. Table 4.1 summarises the performance of the different cost functions and confirm the superior performance of the differential entropy criteria. Finally, figure 4.12 illustrates the plots obtained with Expt(c) and Expt(d) in terms of the number of iterations to convergence of the dual cost function, the convergence error, as well as the profiles of the keyframe selection criteria. The MAD is an increasing function with respect to the intensity error between the registered keyframe and the current

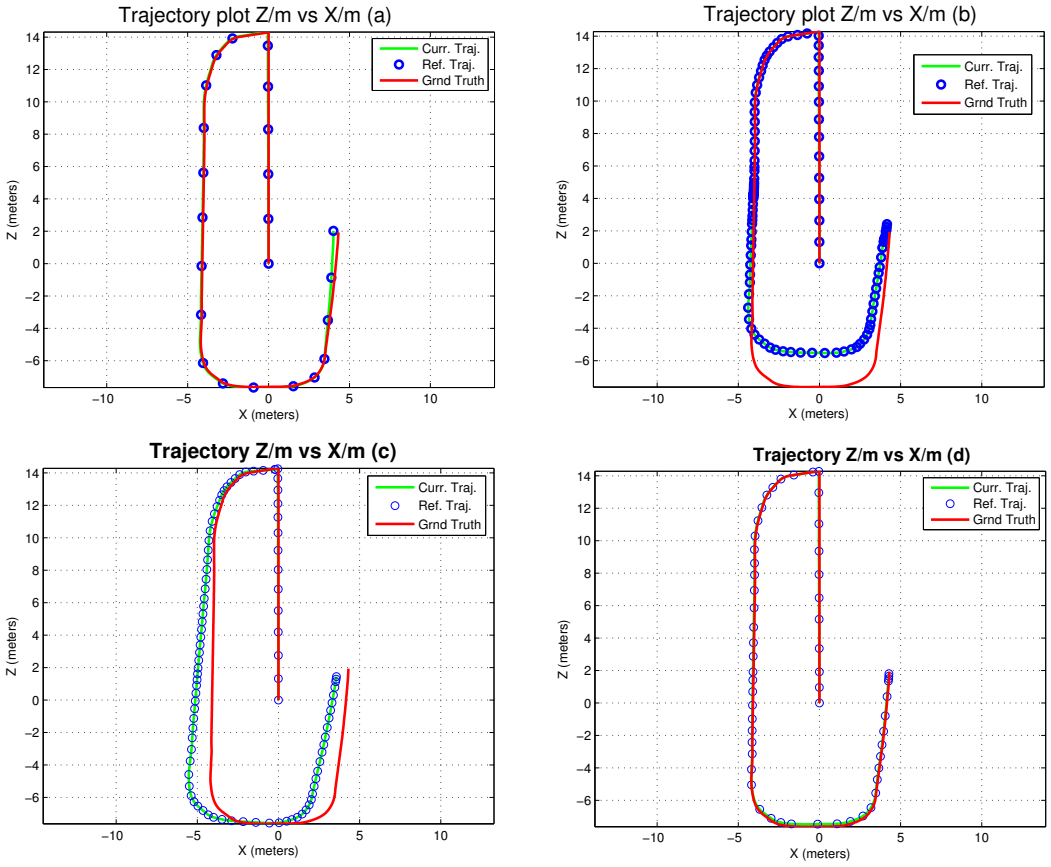

Figure 4.11: *Trajectory comparison generated with different sets of experiments*

frame increases while α decreases when the uncertainty in the pose increases.

4.7.2 Inria Semir dataset

Our experimentations are performed on a dataset of around 170 intensity and depth images within an office environment constituted of several rooms and corridors. During the initial conception phase of the sensor,acquisition was not automatic and thus required a user in the loop to register a snapshot. The sensor was embarked on a trolley and driven around the hallway of Semir building. This acquisition campaign was performed in a stop and go fashion and therefore, framerate varies along with the motion of the user. Figure 4.18 illustrates the various places observed along the trajectory.

Figure 4.15(a) shows the trajectory obtained using the cost functions of equations (4.45) and (4.58) respectively. We observe that the algorithm using vision-only performs poorly in low-textured regions such as corridors or in the presence of reflections from window panes. Such circumstances lead to erroneous estimated poses coming from poor convergence of the algorithm. On the other hand, the *photometry + geometry* cost function takes care of

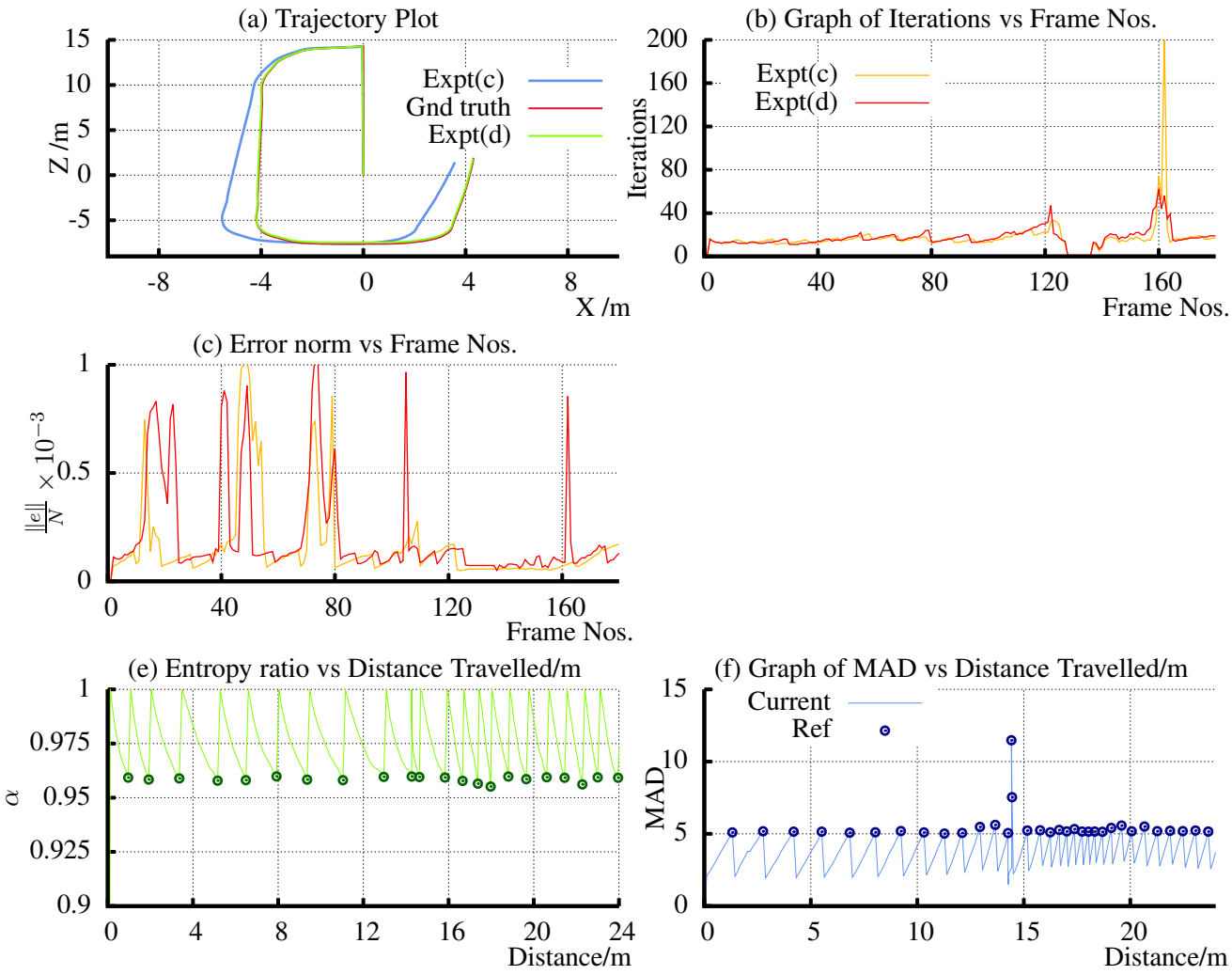

Figure 4.12: *Performance comparison with augmented cost function using the MAD and the differential entropy criteria for Keyframe selection*

these discrepancies by relying more on the depth information available. This is justified by the overall faster convergence of the algorithm as profiled in figure 4.15(b). The high spikes of the figure capped at 200 iterations are due to the non-convergence of the intensity cost function while the new approach still manage to converge at lower iterations. Finally, figures 4.15(c) and 4.15(d), depicts the error norm of each frame at convergence.

Figure 4.16 focusses on the keyframe criteria discussed in section 4.5 for the same dataset: MAD (*method 1*) and Entropy ratio (*method 2*). Over here, we fix the *photometry + geometry* approach with the keyframe criteria as the only variants. While the MAD acts on the residual warping error after convergence, the entropy ratio α abstracts the uncertainty in the estimated pose along the trajectory. The number of spheres initialized for *method 1* is around 50 while *method 2* revealed 27 re-initializations. However, we believe that greater reduction is achievable with lesser inter frame acquisition so that the pose esti-

Figure 4.13: *Snapshots of Inria Semir dataset*

mation algorithm is better initialized leading to a faster and more accurate convergence. A heuristic threshold of 15 was chosen for the case of the MAD and 0.96 for that of α.

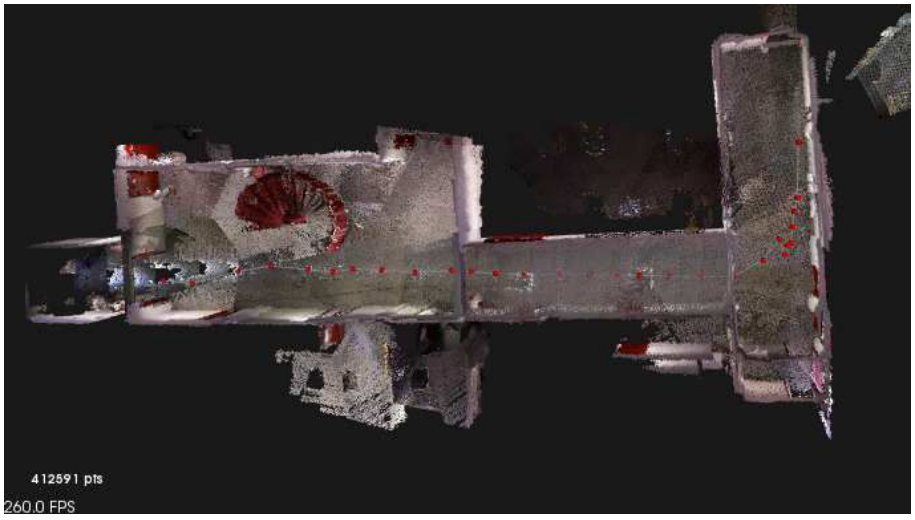

Figure 4.14: *Top view of real trajectory with dual intensity and depth cost function along with entropy ratio α using the Semir dataset*

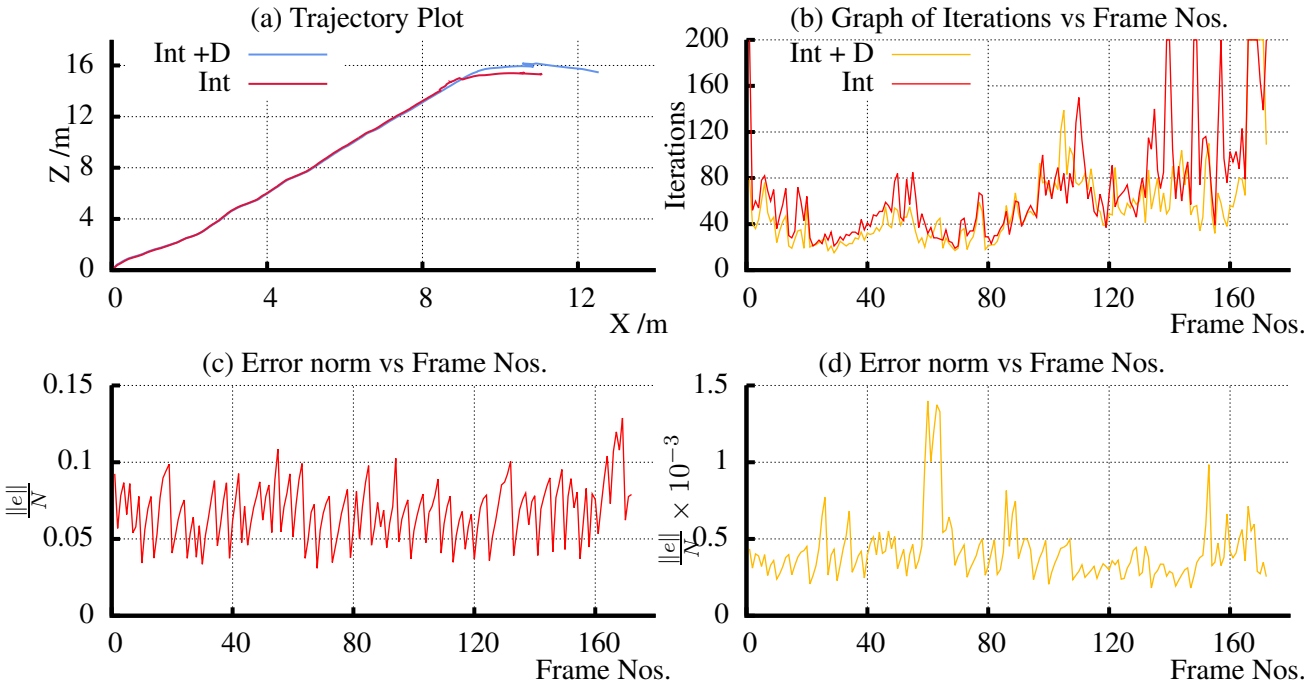

Figure 4.15: *Performance comparison between Intensity only and Intensity and Depth cost functions*

4.7.3 Inria Kahn building dataset (ground floor)

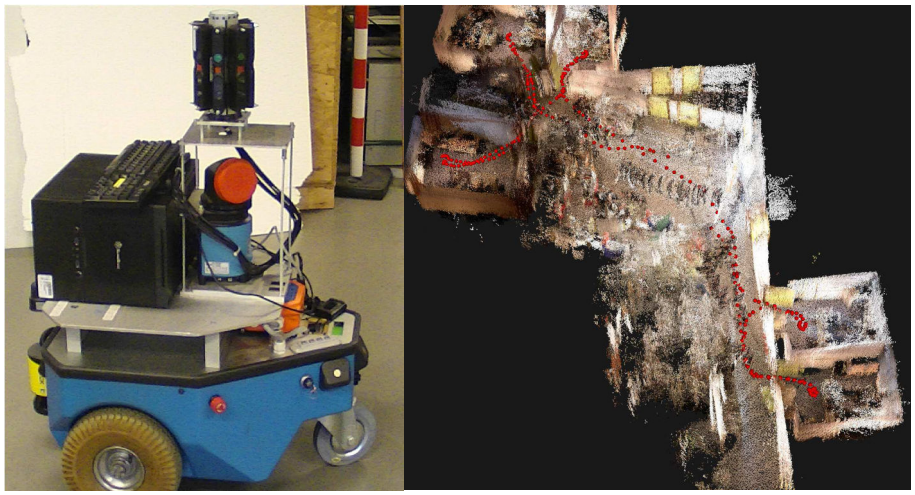

Figure 4.17: *Inria Kahn building experimentation and mapping results*

The spherical sensor is embarked on a mobile experimental platform as shown in figure 4.17 and driven around in an indoor office building environment for a first learning phase whilst spherical RGBD data is acquired online and registered in a database. In this work, we do not address the problem of the *real time aspect* of autonomous navigation and map-

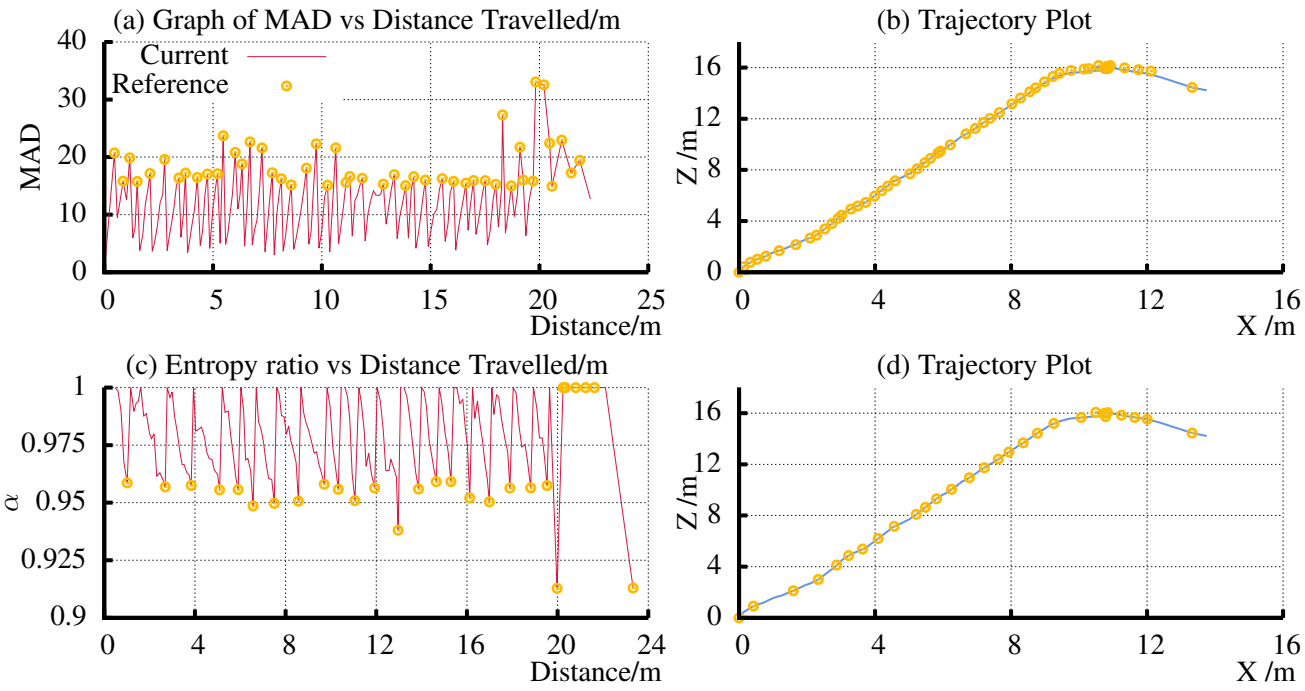

Figure 4.16: *Comparison between MAD vs entropy ratio α using the same augmented photometry + geometry cost function*

Figure 4.18: *Snapshots of Inria Kahn building dataset*

ping but rather investigate ways of building robust and compact environment representations by capturing the observability and dynamics of the latter. Along this streamline, reliable estimates coming from sensor data based on 3D geometrical structure are combined together to serve as a useful purpose for later navigation and localisation tasks. In this experiment, part of a dataset of around 2500 images were used accounting for a total trajectory of around 60 metres. Figure 4.17(right) illustrates the mapping results obtained from the acquired dataset. Though no ground truth was available, the accumulation of

trajectory drift is evident from the misalignment and duplication of wall structures. The environment consists of offices and an open space cluttered with lab equipments as well as other experimental robotics platform. The dots in red show the driven trajectory of the robot along the hallway and into the offices. In the following subsection , we shall tackle the problem of drift using this dataset and find ways to suppress its effect.

4.7.3.1 Metric loop closure

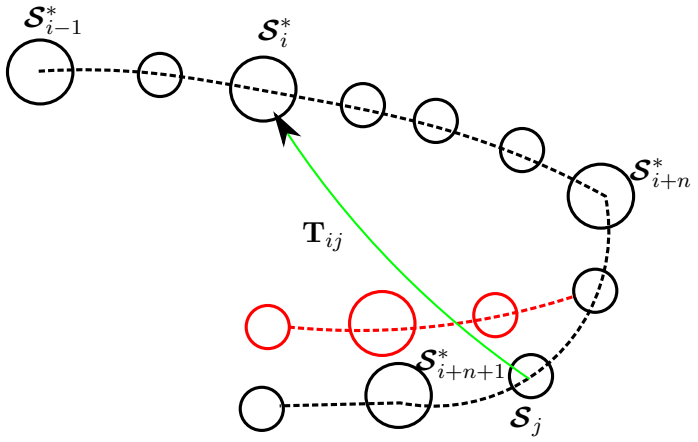

Figure 4.19: *Pose graph correction*

The aim of metric loop closure is to identify when the robot runs into a previously visited area in order to correct the trajectory due to accumulated drift. Given a graph of poses as illustrated in figure 4.19, loop closure is identified by performing a simple euclidean metric check around a radius of the current sphere with all the reference spheres of the graph. This approach is chosen since the area of operation is restricted to an indoor environment and we presume that VO is sufficiently accurate. A simple pose composition between the current sphere $\boldsymbol{\mathcal{S}}_j$ and that of its closely related reference, say $\boldsymbol{\mathcal{S}}^*_i$ is computed a follows:

$$T_{ji} = \ominus T_i \oplus T_j, \tag{4.73}$$

from where, the euclidean distance is obtained as $t_{ij} = ||\mathbf{e}^\top T_{ij}||_2$, where $\mathbf{e}^\top$ is a row vector extracting the translational components. Eventually, a threshold is applied on t_{ij}.

After the identification phase, the next step is to register the pose between $\boldsymbol{\mathcal{S}}^*_i$ and $\boldsymbol{\mathcal{S}}_j$. This obviously can be easily achieved using the registration technique presented in this chapter. However, for dense VO to converge, a good initialisation is required. To proceed, we use a technique often applied with laser based approaches often when tracking is lost. This involves finding an approximate rotation matrix between two measurement couples. In order to apply this to our augmented spherical set $\{\boldsymbol{\mathcal{S}}^*_i, \boldsymbol{\mathcal{S}}_j\}$, several arrays from both intensity and depth information are extracted and the rotation is found using a shifted sum of squared differences (SSD) ,rotated around 2π on the y axis since we assume that the robot is navigating on a horizontal plane $x - z$. The SSD formulation is given as follows:

$$\theta = \underset{\omega}{\operatorname{argmin}} \sum_{k=1}^{\mathcal{L}} \sum_{\omega=0}^{2\pi} \left(\boldsymbol{\mathcal{S}}_{i_k} - \boldsymbol{\mathcal{S}}_{j_k}(\omega)\right)^2,$$

where, $\boldsymbol{\mathcal{S}}_j$ is rotated 2π and $\mathcal{L}$ is the number of arrays used for $\{\boldsymbol{\mathcal{S}}_i^*, \boldsymbol{\mathcal{S}}_j\}$. Figure 4.20 shows an example where two nodes are identified for loop closure and are actually shifted by an unknown angle θ. The last image in the figure gives an example of the outcome of the shifting effect of node 1342 with respect to node 63. Figure 4.21 illustrates how the global minimum of the rotation angle θ is extracted. Eventually, the minimum obtained only from the depth map is used to initialise the transformation between $\boldsymbol{\mathcal{S}}_i$ and $\boldsymbol{\mathcal{S}}_j$ as it was found that θ coming from the depth map was more stable than that computed from intensity images. In order to increase the convergence domain of the registration process, a three level pyramid decomposition was implemented which was sufficient for pose estimation. When the transformation between $\{\boldsymbol{\mathcal{S}}_i^*, \boldsymbol{\mathcal{S}}_j\}$ is recovered, $\boldsymbol{\mathcal{S}}_j$ is referenced back to the "global" reference by the following composition:

$$T_j^* = \oplus Ti \ominus T_{ij}^{corr} \tag{4.74}$$

The "global" reference frame is normally taken to be the first frame acquired when the robot starts navigation. Consequently, the new pose between $\boldsymbol{\mathcal{S}}_j$ and $\boldsymbol{\mathcal{S}}_{i+n}^*$ is also computed using simple pose composition and is thereafter injected into the successive registration process with $\boldsymbol{\mathcal{S}}_{j+1}$ and $\boldsymbol{\mathcal{S}}_{i+n}^*$ to rectify the trajectory. A summary of the algorithm is provided below:

Algorithme 1 Loop Closure and Trajectory Correction

Require: Spherical RBGD VO
  Check for Loop Closure (LC)
  if LC = true **then**
    for each pyramid level **do**
      Compute semi-dense VO btw node i and node j
      move to next level
    if VO converged **then**
      Apply trajectory correction
    else
      No correction applied

Finally, figure 4.22 depicts the reconstruction quality obtained with the technique of loop closure and that of the original reconstructed map obtained only with VO. It is observed that though the misalignment between the wall structures have been reduced (green square), other artifacts have been introduced in the map (red circle). The reason put forward can be obviously a wrong pose initialisation that led to an erroneous registration process. One loophole of this method is that the trajectory prior to $\boldsymbol{\mathcal{S}}_j$ remains unchanged and also, the uncertainty of the poses have not been taken into account. Therefore, a better alternative would rather be to perform pose graph optimization by taking into consideration all the poses between $\boldsymbol{\mathcal{S}}_i$ and $\boldsymbol{\mathcal{S}}_j$ as well as propagating their uncertainties across the whole chain.

Figure 4.20: *Top and centre: Illustration of loop closure between nodes 63 and 1342. Bottom: recovered rotation using outlined SSD technique.*

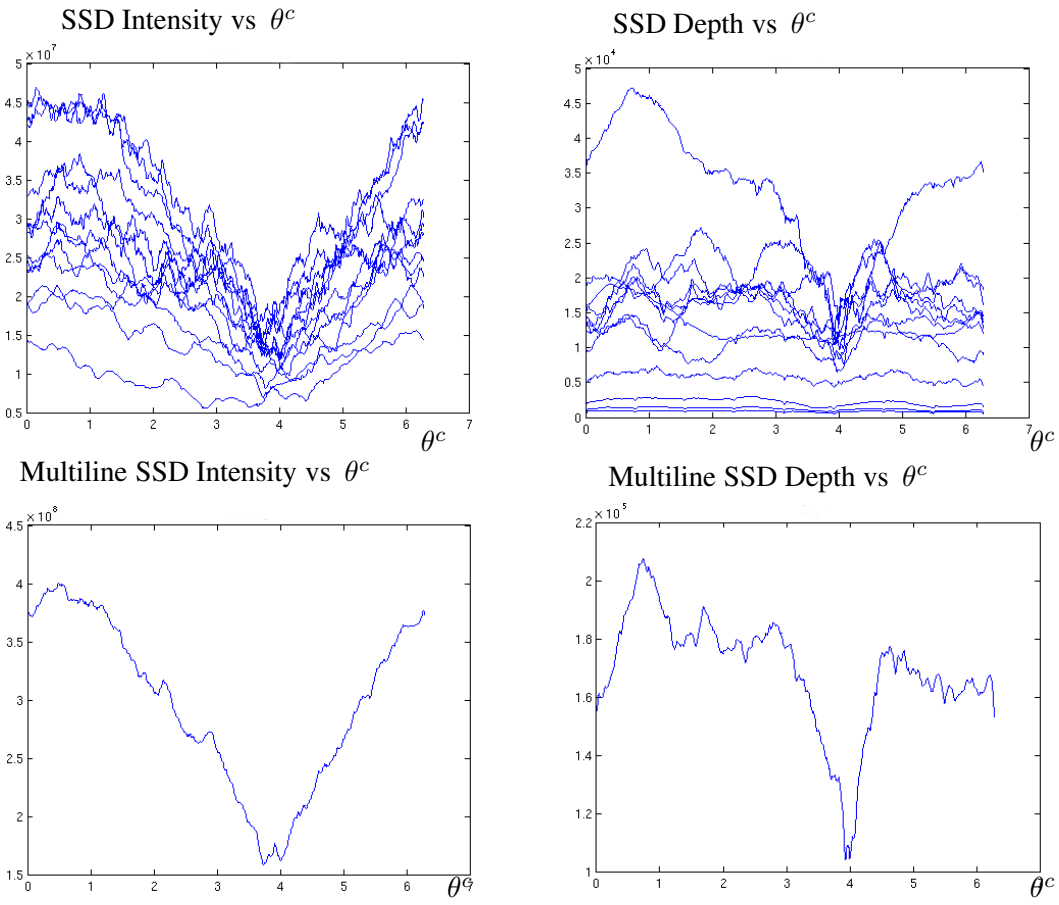

Figure 4.21: *Rotation estimation using an SSD cost function*

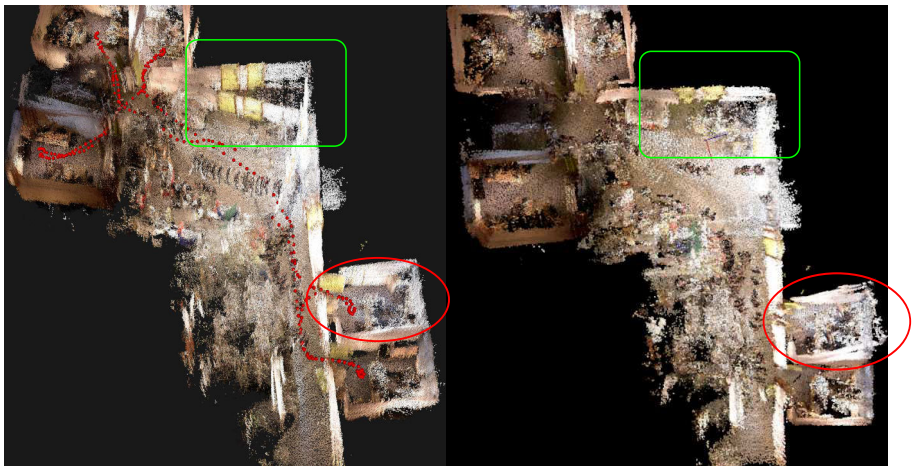

Figure 4.22: *Reconstruction comparison with metric loop closure using Kahn dataset*

4.7.4 Results with Garbejaire Dataset

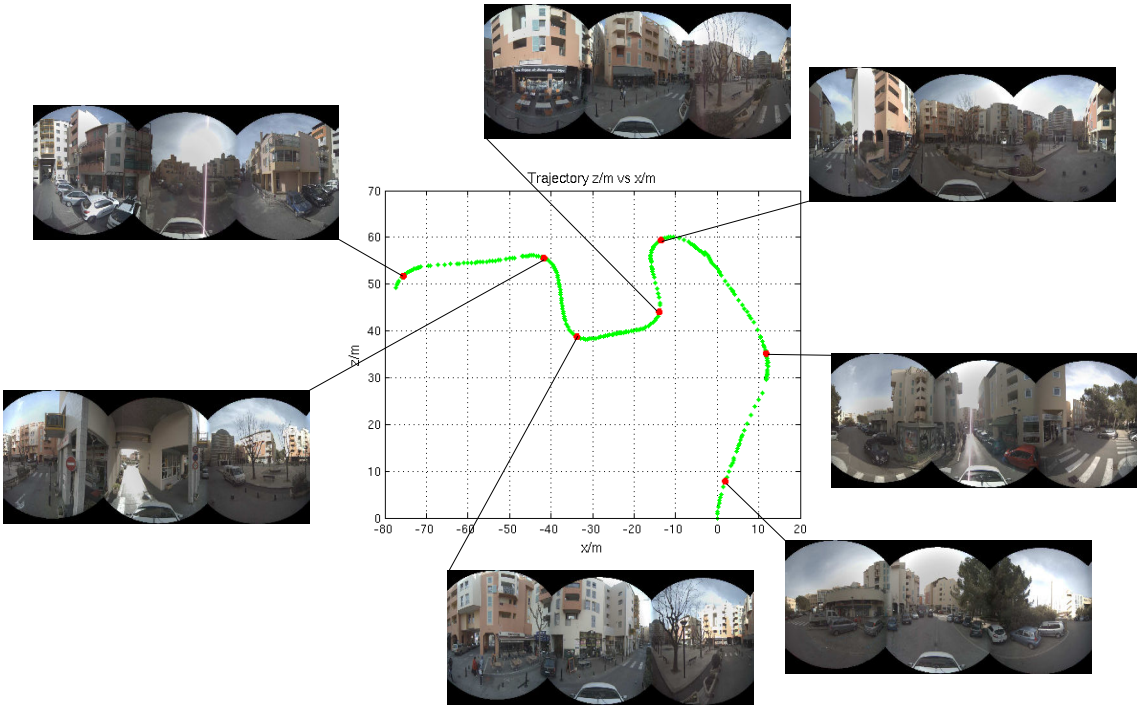

Figure 4.23: *Trajectory reconstruction with Garbejaire dataset*

This urban dataset consists of more than 20K images using our outdoor spherical sensor mounted on a car gallery as depicted in figures 1,3.12. Spherical RGBD images are reconstructed offline. Our algorithm is applied only on a specific portion of the total travelled trajectory as indicated in the snapshot of google's map viewer. It is observed that VO degenerates on this particular dataset as illustrated in figure 4.24(a, red). The only reason

Figure 4.24: *Degenerate conditions and VO failures on full trajectory reconstruction*

evoked is the extremely noisy depth maps built from SGBM [Hirschmuller 2006]. To get VO working on this measurement set, we resort to depth map filtering which consists of reprojecting and fusing depth maps in a window of 5 frames with its centre chosen as the reprojection frame. This fusion technique is not part of our work and is further elaborated in [Martins 2015]. Re-application of VO with the filtered depth maps results in a more consistent trajectory as shown in figures 4.24(a, blue),(c). This particular trajectory consists of around 3700 augmented spheres out of which only 850 keyframes are recorded. Analysing figures 4.24(c) and (d), it is evident that the loop is not closed as desired. This is due of erroneous pose estimates along the trajectory which are further propagated down the chain. These wrong (rotational) estimates occur mainly in regions where the vehicle is negociating curbs, resulting in considerable changes across viewpoints leading to failures of the direct method. It would therefore be interesting to detect when VO failures occur in the

optimisation loop in order to anticipate these discrepancies. Though the global trajectory is not similar to the real one, it is observed that locally, the trajectory is quite "piecewise" consistent. Figure 4.24(b) depicts a snapshot our Opengl viewer. Figure 4.23 shows part of the trajectory with variations in observations registered along the curbs.

4.8 Conclusion

This chapter presents a robust direct semi/dense visual odometry technique using a hybrid photometric and geometric cost function in order to overcome the shortcomings of intensity only based pose estimation techniques such as illumination or features' mismatching in poorly textured areas. A new criteria based on differential entropy overuns the previous MAD criteria for keyframe selection. The advantages are two-fold. Firstly, the explored environment is represented by less keyframes, hence a more compact pose graph representation is achieved. Secondly, using less keyframes resorts to reduced error integration due to frame to keyframe registration, hence helps in the reduction of the overall tracking drift as shown with the results of the synthetic dataset. Our algorithm was further evaluated on two real datasets acquired under two different conditions in two different buildings. The smaller *Inria Semir* dataset present a scenario where the camera moves through a hallway made up of textureless corridor surfaces and big window frames. The *Inria Kahn building* dataset is more elaborate, consisting of office spaces as well as a cluttered area "dumped" with lab equipments. While our algorithm was fairly tested on the first dataset, the second one exposes its weakness whereby acute manoeuvres with the robot lead to erroneous pose estimation and tracking drift.

In order to address these problems, a simple metric loop closure algorithm has been implemented at local graph level. Over here as well, it is observed that registration at two different viewpoints requires a good initialisation for the cost function to converge to the global minimum. Though the overall reconstruction quality has been fairly improved, other artifacts have been introduced in the map. It is deduced that this discrepancy might be coming from the multi modal minima of the SSD function which outputs a angular rotation which is then used to bootstrap our optimisation. The best way to tackle this problem would be to first perform loop closures at an appearance based level as in [Chapoulie *et al.* 2011] followed by local optimisation techniques for pose correction as in [Kümmerle *et al.* 2011].

In the next chapter, we change our approach and focus on ways of how to improve the information content of the augmented sphere. It shall be seen that, by applying a filtering technique on both geometry and photometry, better results are obtained.

CHAPTER 5

Towards Accurate and Consistent Dense 3D Mapping

5.1 Introduction

Nowadays, mobile robots are expected to be fully autonomous while exploring and interacting with their immediate surroundings. In fact, the kind of environment under which the robot is expected to operate is very much unstructured and dynamic. Ephemeral subjects such as people moving around simply distract an otherwise static map. Furthermore, robots should deal with perceptual aliasing, weather changes, occlusions or illumination variations. Therefore, changes are really unpredictable and occur at different rates; they can be abrupt, gradual, permanent or non-permanent. Hence, the capability to dissect these occurrences is very much desirable. Moreover, the system must have the intelligence to reflect on its old state and be able to revert back when changes are just temporary. For slow and gradual changes, *e.g.* seasonal changes, construction buildings, present day to day challenges for *servicebots* to constantly adapt to these inexorable changes. Therefore, the concept of lifelong mapping is implemented in these systems so that the robot is able to constantly repair and increment its map building capability to be able to maintain a valid environment representation over a long period of time. The main challenge of mapping dynamic environments comes from the fact that the environment model can change in unpredictable ways. Hence, the internal representation of the map in the mobile robot can become easily out of date leading to catastrophic effects on the performance and efficiency of the planning and navigation tasks. In this context, we devise a method interleaved between SLAM and computer graphics community in order to track and filter out dynamic 3D points as well as update the static part of the map along a driven trajectory.

5.2 Methodology

Our aim is concentrated around building ego-centric topometric maps represented as a graph of keyframes, spread by spherical RGB-D nodes. A locally defined geo-referenced frame is taken as an initial model which is then refined over the course of the trajectory by neighbouring frame rendering. This not only reduces data redundancy but also helps to suppress sensor noise whilst contributing significantly in reducing the effect of drift. A first

approach is devised where the depth signal of a 3D point is reconstructed over time and a statistical analysis is made in order to detect its inconsistency. Thereafter, accumulation of depth values making up the signal are fused only if the consistency of that 3D point is maintained over an explored trajectory.

The second approach involves a generic uncertainty propagation model leaned upon a data association framework for discrepancy detection between the measured and observed data. We build upon the above two concepts to introduce the notion of landmark stability along the trajectory. This is an important aspect for intelligent 3D points selection which serve as better potential candidates for subsequent inter frame to keyframe motion estimation. A fusion stage follows by considering both photometric and geometric information in order to update consistent data over the explored trajectory.

5.3 A first approach to depth map fusion

The first fold of this chapter is based upon the tracking of a 3D point by continuously projecting the current depth map obtained along the trajectory in its annotated reference frame obtained from the Keyframe criteria discussed in the previous chapter. Rasterising the depth information in a common frame allows making the profiling of the depth information and hence detect possible ruptures in the signal resulting from noise or occlusion phenomena. The signal rupture model is based on a statistical event-based test, namely the Page-Hinckley Test (PH-T). In this section, we shall on a first front, undermine the working principle of the test and thereafter explain how we proceed to use the information provided to fuse the depth map.

5.3.1 Depth inconsistency detection

Page-Hinckley Test(PH-T) is a statistical event detection test used in data mining and statistics for large amount data treatment. Data related to spatio-temporal processes are often streamed in the form of time-bounded numerical series. Therefore, an important task in exploration of time related data is the detection of abrupt changes or fluctuations of the studied variable over time. In this work, we shall apply P-HT to detect inconsistency in our back-projected data to decide on the stability of a certain depth value before the ultimate data fusion stage (section 5.3.2). Inconsistencies may be cased by data noise or occlusion occurrences identified as disruptions in the normal signal flow.

P-HT as defined in [Andrienko *et al.* 2010] is designed to monitor drifts in the mean of a time series and is given by:

$$m_T := \sum_{t=t_0}^{T} (\rho_t - \bar{\rho}_T - \delta), \tag{5.1}$$

where, $\bar{\rho}_T$ is the empirical mean of the projected depth values from t_0 to T. At each

step of the mean drift test, a variable $M_T = min(m_{t_0}, ..., m_T), \forall t$ is evaluated and an alarm is raised whenever a certain threshold on the difference between m_t and M_T is met, i.e $m_T - M_T > \beta$. Both β and δ above are heuristically tuned parameters based on a desirable step tolerance.

Algorithme 2 Page-Hinckley Test

Require: Time series ρ_k of length n
Require: Parameters β, δ
  return event
  Initialise $t_0 \longleftarrow 1$, $acc \longleftarrow 0$, $min \longleftarrow \infty$
  for $T = 1$ **to** n **do**
    $m_T \leftarrow 0$
    $acc \leftarrow acc + x_T$
    $\hat{\rho}_T \leftarrow acc/(T - t_0 - 1)$
    for $K = t_0$ **to** T **do**
      $m_T \longleftarrow m_T + (\rho_K - \hat{\rho}_T - \delta)$
    if $min > m_T$ **then**
      $min \longleftarrow m_T$
    if $m_T - min > \beta$ **then**
      report event $(T, m_T - min)$
      $t_0 \leftarrow T$
      $acc \leftarrow 0$
      $min \leftarrow \infty$

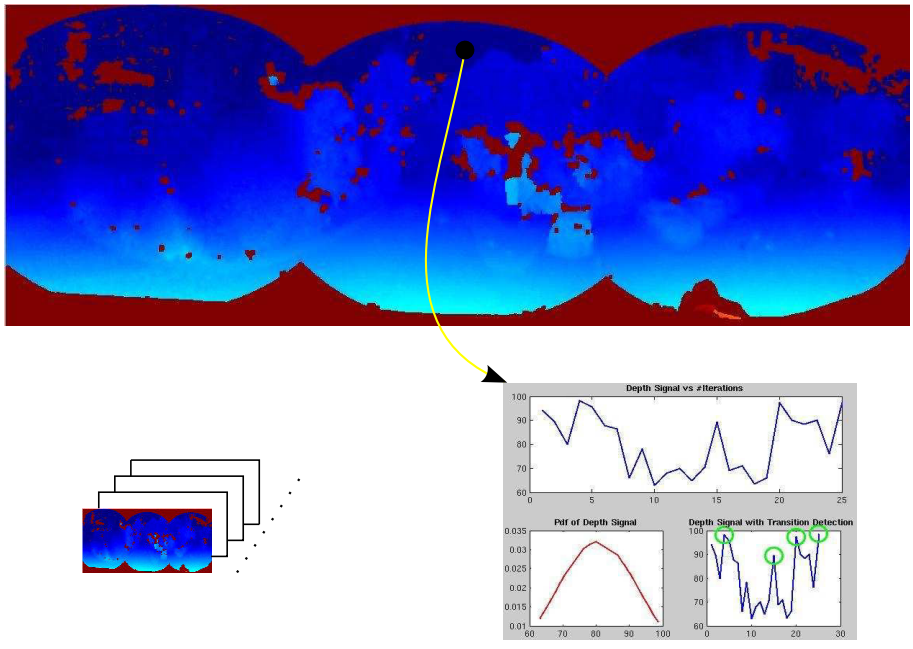

Figure 5.1: *Page-Hinckley Test: events detected*

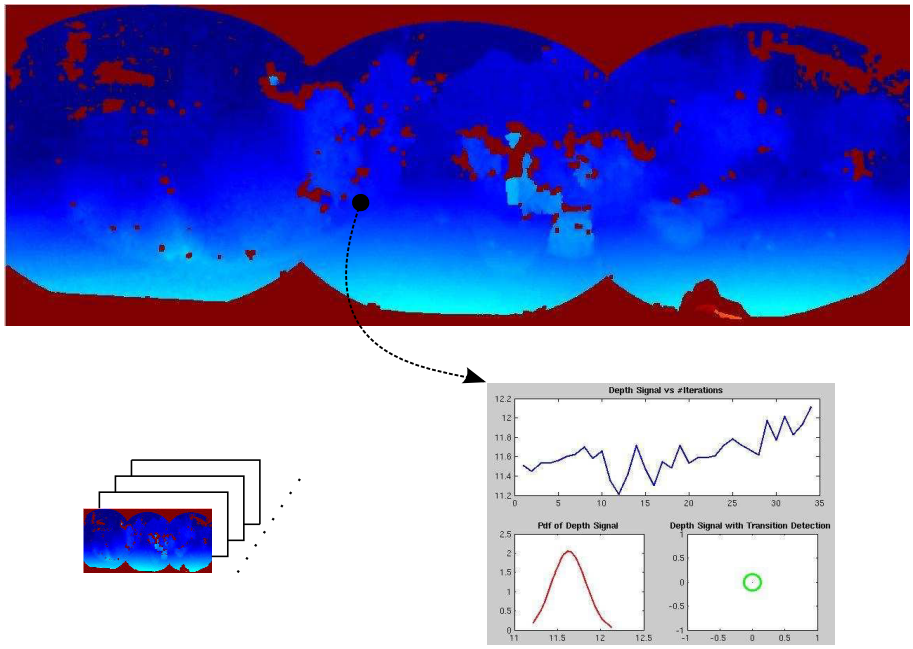

Figure 5.2: *Page-Hinckley Test: no events detected*

Figure 5.1 above illustrates the reference depth map discretised on a spherical grid map, $\mathcal{D}^* \in \mathbb{R}^{m \times n}$, with vertices, $\mathcal{V}_i^*(\theta, \phi), \forall i \in$ (m,n). When new current frames are acquired, they are rasterised in the reference view and stacked parallelly, with each stack representing a depth map hypothesis. Accumulating n hypotheses results in a depth profile for each vertex. The figure picturises a scenario where a point belonging to a far away point is randomly picked up. This point may be coming from the sky entity for example. It is observed that this point is very noisy and the PH-T triggers various alarms to prompt the inconsistency of the signal. The noisy signal gives rise to a rather flat probability distribution function (pdf) with a large variance.

On the other hand, figure 5.2 shows the statistics related to a point picked from a scene closer to the camera, around 11 metres. The signal shows far better consistency with no alarm raised and exhibits a sharper pdf. Therefore, the hypothesis that faraway points are much noisier than close up points is confirmed as the uncertainty of a depth value varies quadratically to the corresponding z-distance. This uncertainty model will be later used for the fusion step.

5.3.2 Inverse Warping and Depth Map Fusion

Our approach to metric map building is an ego-centric representation operating locally to sensor data. Therefore, we refrain from using a global model for fusion and instead, a locally defined reference frame is chosen where the stages of inverse warping and depth map fusion will take place.

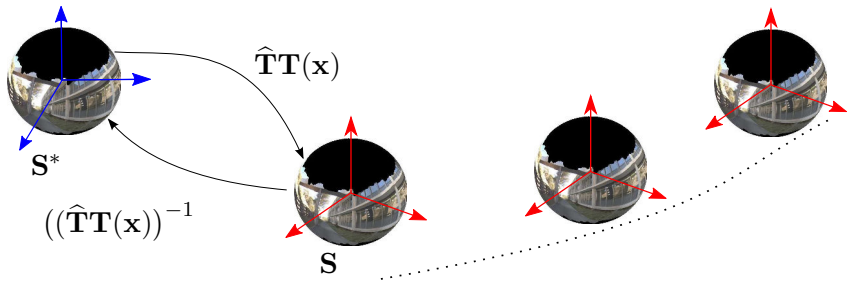

Figure 5.3: *Inverse warping on reference frame*

After frame to Keyframe spherical visual odometry (section 4.4), the current depth map of $\mathcal{S}_n$ is rendered in the reference frame of $\mathcal{S}^*$ as illustrated in figure 5.3. An inverse warping operation is performed and the resulting non-uniform mesh grid is interpolated on a regular spherical grid of the reference frame to attribute depth values to each of its vertices $\mathcal{V}_i(\theta, \phi)$. The steps described above are repeated for a set of augmented spheres $\mathcal{S}_n$ over a sliding window of n views, e.g $n = 10$. Back projecting and stacking of depth maps in $\mathcal{S}^*$ frame yields a volumetric voxellized structure and a discretised depth signal emerges for every vertex of the spherical grid fused incrementally by using a weighting average filter as follows:

$$\mathcal{D}^*_{k+1}(\mathbf{p}) = \frac{\mathbf{W}^D_k(\mathbf{p})\mathcal{D}^*_k(\mathbf{p}) + \Pi_D(\mathbf{p})\mathcal{D}_w(\mathbf{p})}{\mathbf{W}^D_k(\mathbf{p}) + \Pi_D(\mathbf{p})}, \tag{5.2}$$

where, $\mathcal{D}^*_k(\mathbf{p})$ and $\mathcal{D}_w(\mathbf{p})$ are the reference and the warped depth map respectively. The weight for each depth entity is obtained using the uncertainty model developed in [Khoshelham & Elberink 2012] and applied as follows:

$$\Pi_D = \frac{1}{\sigma^i_\rho} : \sigma^i_\rho = \frac{\rho^2 m \sigma_d}{fb}, \tag{5.3}$$

where, f:focal length, b: baseline, ρ: depth, σ_d: disparity uncertainty and m: constant with weight update matrix leading to :

$$\mathbf{W}^D_{k+1} = \mathbf{W}^D_k + \Pi_D \tag{5.4}$$

Concurrently, an indexed map $\mathcal{M}^*(\mathcal{V}(\theta_i, \phi_i))$ is generated based on the outcome of P-HT. Whenever a hypothesised back projected depth value $\rho^*_n(\theta_i, \phi_i)$ gives a positive response to the test, its corresponding status counter in $\mathcal{M}^*$ is incremented. This gives an indication of the stability of the point with respect to data noise or occlusions. The greater the count number registered, the higher the confidence level placed on that particular point. In this way, points associated with the highest score are marked as stable as they have been perceived all the way along the trajectory of the n view tuple, while unstable points are labelled as outliers in the fused depth map.

Finally, we build a saliency map based on 4.2.4, excluding outliers mentioned above. The saliency map is the result of careful selection of the most informative pixel in descending order based on the photometric Jacobian matrix obtained from equation (4.58). Instead of naively using the entire dense information of the depth map, the computational burden is relaxed by using a subset of the top 10%-20% extracted for the process of Spherical registration.

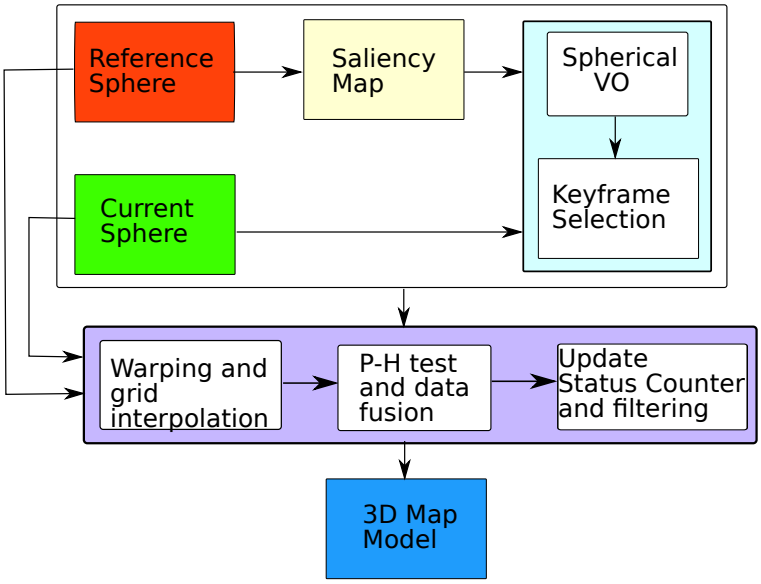

Figure 5.4: *Pipeline using Page-Hinckley Test*

5.3.3 Results

Figure 5.5: *Reconstruction comparison with Inria Kahn dataset. Left: No fusion ,Right: fusion with algorithm 2*

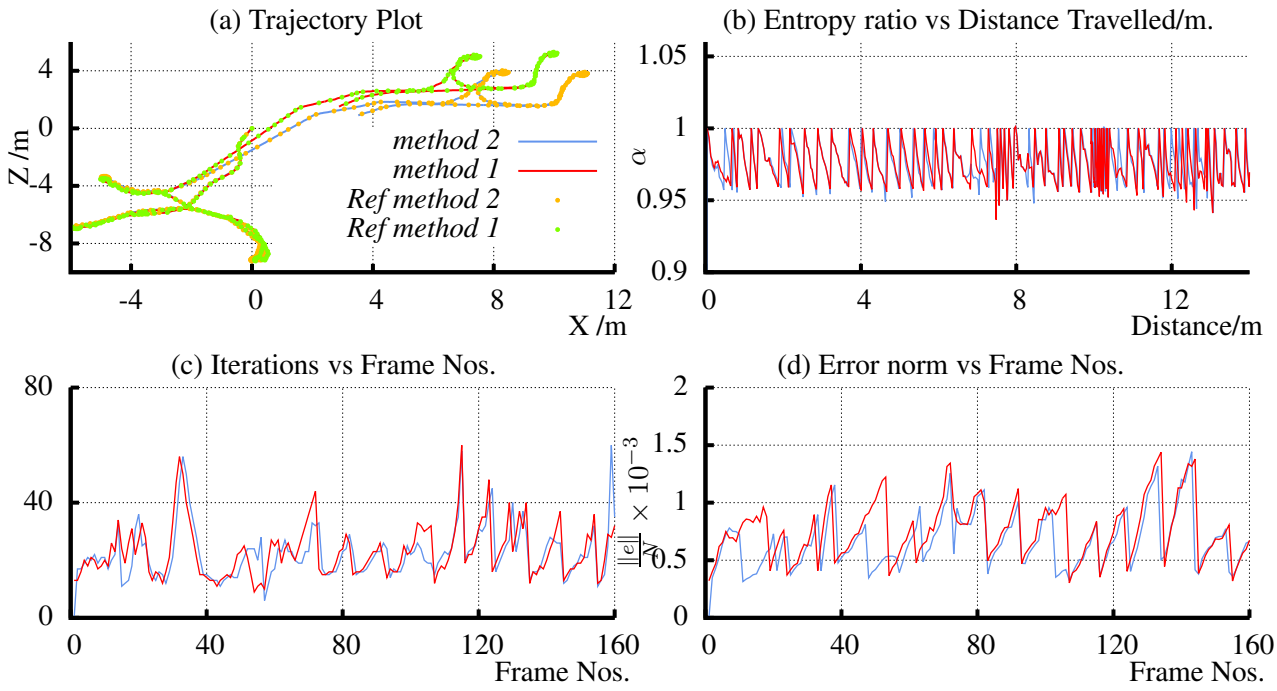

Figure 5.6: *Performance comparison with Kahn0 dataset*

Our algorithm has been extensively tested on the *kahn building* dataset introduced in section 4.7.3. The aim is to evaluate the approached methodology using PH-T pipeline. Two methods have been compared as follows:

- Original RGB-D registration (***method 1***)
- RGB-D registration with depth map fusion using PH-T (***method 2***)

Figure 5.5(right) & (left) demonstrate the point cloud reconstruction obtained from two experiments; (***method 1***) and (***method 2***). In detail, reconstruction with *method 1* demonstrates the effects of duplicated structures (especially surrounding wall structures) which is explained by the fact that point clouds are not perfectly stitched together on revisited areas due to inherent presence of rotational drift, which is more pronounced than translational drift. However, these effects are reduced by the fusion stage but not completely eliminated. The red dots on the reconstruction images are attributed to the reference spheres initialised along the trajectories using the keyframe criteria described in section 4.5.2. 270 initialisations were recorded for *method 1* while 252 key spheres were registered for *method 2*.

Finally, figure 5.6(a) illustrates the total trajectory travelled in the building with the reference spheres for *method 1* and *method 2* (in green and orange respectively). Figure 5.6(b) depicts the behaviour of our keyframe selection entropy-based criteria α (with a chosen threshold of 0.96). Figure 5.6 (c) and (d) show the convergence rate of the registration

stage (average of 20 iterations per frame to keyframe alignment), and the error norm at convergence for both methods respectively.

To conclude, while the PH-T pipeline results in slightly better global reconstruction of the scene, it does not contribute much to this fusion technique. Moreover, the number of reference spheres have been reduced by a slender margin of 6.7% and hence does not bring much to the trajectory correction since error drift occurring from frame to Keyframe odometry has not been considerably reduced as expected. Furthermore, a simplistic error model is considered for the warped depth map without taking into account the transformations undergone across the chain before being represented in the reference frame , i.e, the uncertainty related to the warping and interpolated phase has been ignored. To further improve the map, a back-end SLAM solution is presented in the next section where the trajectory is corrected using manual loop closures. Eventually, in the second part of this chapter, an improved error and fusion model is presented which takes into account both sensor and pose uncertainties.

5.3.4 Pose graph optimisation

The VO technique introduced in the previous chapter 4 is an integral part of Front-End graph SLAM where the aim is to generate a graph of nodes connected by edges. The relationship between nodes and edges are defined by geometric constraints coming from sensor data interpreted by sensors – in our case perception sensors are then main focus. No matter how robust front-end SLAM tries to be with respect to outliers or wrong initialisations, at some point of time, degenerate conditions are bound to happen due to the high non linearity of the SLAM problem itself. This can be coming from a wrong pose estimation occurring from noisy or erroneous data or insufficient number of inliers to provide a smooth convergent global minimum. To obtain a reliable pose graph, hence a reliable map for robot navigation, the SLAM Back-End comes into play. It's aim is to find the best configuration of nodes that minimise the error induced by edges from the front-end.

Given a state vector $\mathbf{x} = (\mathbf{x}_1, \cdots\cdots, \mathbf{x}_n)^\top$ where the variable $\mathbf{x}_i$ is the pose of node i related to a possible robot or landmark position. For robots operating in full 6 DOFs, as in our case, pose $\mathbf{x}_i$ is therefore 6D whilst point features /landmarks are represented in 3D. The error function $\mathbf{e}_{ij}(\mathbf{x})$ defined for a single edge between node i and j is given by the difference between the observed measurement $\mathbf{z}_{ij}$ and the expected measurement $\widehat{\mathbf{z}}(\mathbf{x}_j, \mathbf{x}_j)$ for the current state as follows:

$$\mathbf{e}_{ij}(\mathbf{x}) = \widehat{\mathbf{z}}(\mathbf{x}_i, \mathbf{x}_j) - \mathbf{z}_{ij} \tag{5.5}$$

The measurement function $\widehat{\mathbf{z}}$ defined above normally depends on sensor set up. If only pose to pose constraint is used, only transformations between poses are required. On the other hand, for pose to landmark constraint, the reprojection error of the observed landmark into the frame of the observing pose is considered. The cost function encapsulating inter-nodal

constraints is then given by:

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}} \sum_{ij} \mathbf{e}_{ij}(\mathbf{x})^\top \Sigma_{ij} \mathbf{e}_{ij}(\mathbf{x}), \tag{5.6}$$

which is classically solved using the unconstrained optimization technique defined in section 4.2.1.1. However, one particularity about the underlying structure of the Jacobian and the Hessian matrices is that they are sparse due to links established between the nodes or observability conditions of landmark i in node j for example. Solving a huge system of equation with sparse matrices is memory and computation inefficient. In literature, such systems are resolved using sparse matrix decomposition such as Cholesky or QR methods. Levenberg-Marquardt, Powell's Dog leg, Stochastic Gradient descent and its variants are alternative approaches for computational or convergence enhancements. For the case of a strictly convex problem, all the above-mentioned methods should converge to the same global minimum. However, due to the non-linear nature of the measurement, hence the non linear formulation of the SLAM problem, a global minimum is not always guaranteed. Recently, two robust solutions are presented in the work of [Agarwal 2015], namely the Max Mixture and the Dynamic Covariance Scaling approach.

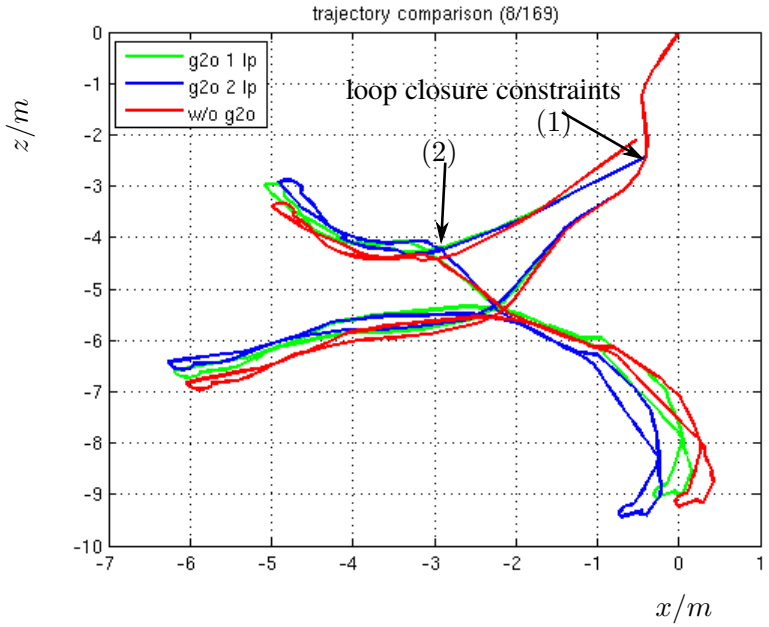

Figure 5.7: *Trajectories obtained with pose graph optimisation*

Many software solutions have been proposed in literature, to name a few: TORO of [Grisetti *et al.* 2007], the sparse bundle adjustment library of [Konolige 2010], the g2o library of [Kümmerle *et al.* 2011]. In this work, g2o is the preferred choice due to its recently improved implementation based on TORO and has been currently used for RGB-D SLAM pose graph optimisation. To correct the trajectory resulting from an erroneous pose estimation or from accumulated drift, a loop closure is manually inserted in the graph and

the initial pose graph obtained from the front-end SLAM is injected into g2o to produce a graph of optimised poses. Figure 5.7 shows the initial trajectory compared with the trajectory output from g2o using 1 and 2 loop closures applied sequentially. The results are better inspected using point cloud reconstruction as displayed in figure 5.8. Figure 5.8(a) depicts the original reconstruction without fusion (chapter 4), Figure 5.8(b), shows reconstruction with the fusion technique as discussed in the previous section 5.3.2. Figure 5.8(d) and (c) show reconstruction with the fusion technique and optimised using 1 and 2 manually inserted loop closures respectively. It is observed that, using a single loop closure (1) helps to improve the graph, hence the reconstruction while a second loop closure (2) added does not contribute much to the reconstruction quality as expected. The reason evoked is that a wrong pose estimation has occured most probably at the second loop closure due to a delicate door passing scenario, leading to an erroneous estimated pose. This biased pose may be coming from the information acquired from the depth map due to the range limitations of the sensor which provides no measurement in regions close to the door. With the first loop closure, the optimiser forces the graph by meeting the inserted constraint. It has to be pointed out that in this experiment, roughly half of the nodes of the full graph is used, some 170 reference nodes in order to analyse and anticipate the area where the problem of dead reckoning is most prominent. The idea was to later insert this back-end SLAM module to work concurrently with the front-end in order to detect loop closures on the fly so that the local trajectory is rectified when incoherent pose estimates are obtained. Due to implementation issues, we were not successful in bridging the link between front-end and back-end SLAM but this part remains vital for a complete functional SLAM system and shall be tackled in our future endeavours.

5.4 An improved approach to environment modelling

An important aspect of environment modelling is the intelligent treatment of data in the construction of efficient quality models. The underlying reason concerns inherently noisy measurements induced by sensors. Noise cannot be completely eliminated but rather, their notoriously adverse effects can be suppressed. Three predominant error treatments approaches are :

- Random errors: coming from the internal circuitry of range measuring devices, the estimated measurement is normally assumed to be bounded around the true numerical value. Such estimates are normally modelled with a zero mean random Gaussian noise in order to compensate for the discrepancy.

- Systematic errors (also static): occur from incorrect acquisition of sensor data itself. These sensor readings are often categorized as false positives. Errors can be due to the experimental set up itself where the system calibration errors, for example, have not been properly tackled.

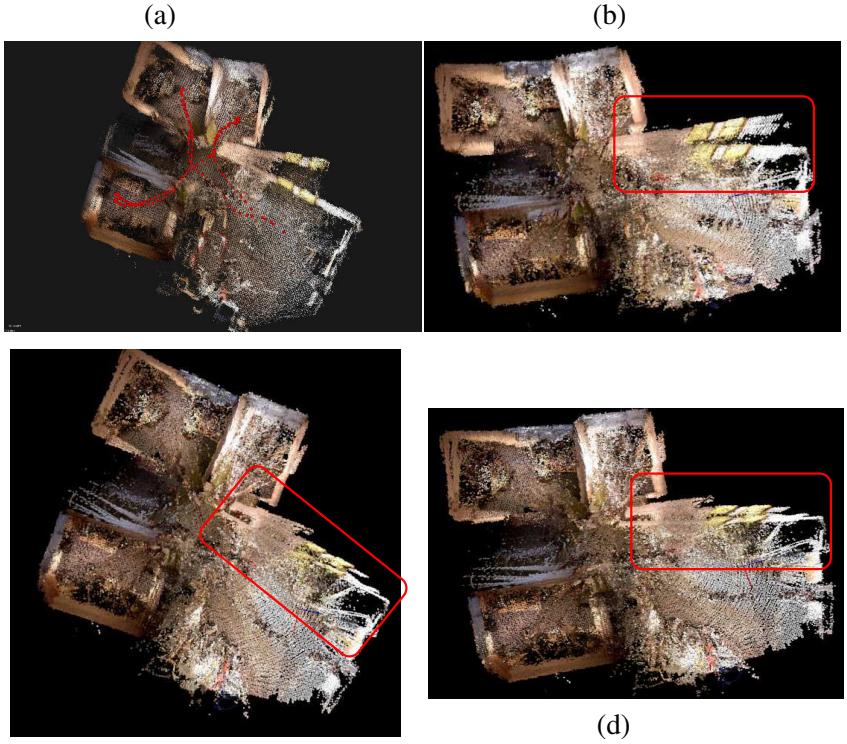

Figure 5.8: *Results with pose graph optimisation*

- Dynamic errors: coming from the measurement uncertainties. Occlusion, disocclusion phenomena fall into this categrory, errors coming from erroneous pose estimation resulting in noisy measurements being integrated in the global map for example. If these errors are not duly anticipated, they result in visual odometry failures.

Therefore, these types of errors must be carefully handled in order to conceive a model as accurate as possible. Incorrect environment models can be detrimental to mobile robot navigation and exploration tasks. A good technique to tackle the above mentioned uncertainties is to accumulate data over time followed by filtering techniques. The work of [Stephen *et al.* 2002] is a good illustration whereby a Triclop stereo vision system (three cameras in a triangular configuration facing the world, giving rise to three stereo pairs). The approach is feature based with least means square minimization to compute visual odometry. A robot odometry model is then fused using an Extended Kalman filter. Landmarks are initialised and propagated using an error model derived from stereo. The main highlight is that a feature database is initialised keeping track of its corresponding landmark location, scale, orientation, feature hits and misses count numbers. This framework is then used to refine landmark positioning over a driven trajectory based on a set of simple heuristics.

The work presented in this section is directly related to two previous syntheses of [Dryanovski *et al.* 2013] and [Meilland & Comport 2013a]. The former differs from our approach in the sense that it is a sparse feature based technique and only consider depth

information propagation. On the other hand, the latter treat a similar dense based method to ours but without considering error models and a simpler Mahalanobis test defines the hypothesis for outlier rejection. Additionally, we build on a previous work of [Meilland *et al.* 2010] which constitutes of building a saliency map for each reference model, by adding the concept of stability of the underlying 3D structure.

Key contributions of this work are outlined as follows:

- development of a generic spherical uncertainty error propagation model, which can be easily adapted to other sensor models (e.g. perspective RGBD cameras)
- a coherent dense outlier rejection and data fusion framework relying on the proposed uncertainty model,yielding more precise spherical keyframes
- dynamic 3D points are tracked along the trajectory and are pruned out by skimming down a saliency map

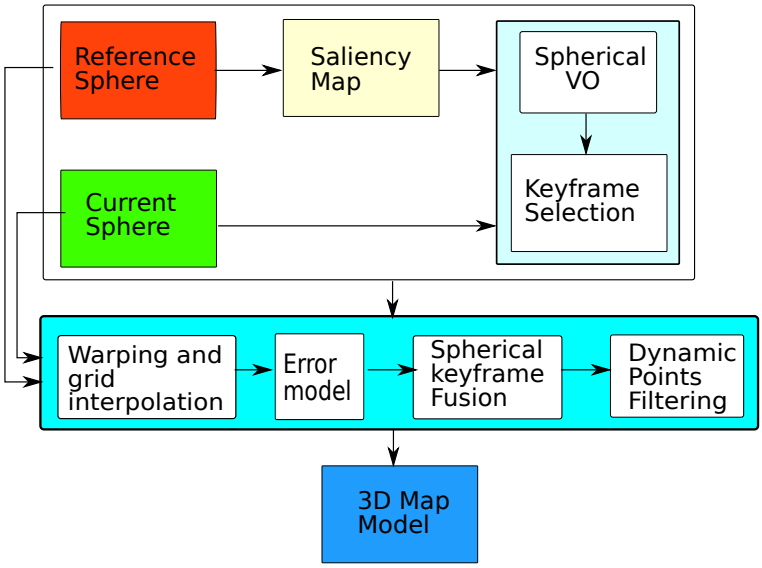

Figure 5.9: *Pipeline using the uncertainty model*

5.4.1 Error modelling and propagation

As mentioned earlier, our approach to topometric map building is an egocentric representation operating locally on sensor data. The concept of proximity used to combine information is evaluated mainly with the entropy similarity criteria after the registration procedure. Instead of performing a complete bundle adjustment along all parameters including poses and structure for the full set of close raw spheres $\mathcal{S}_i$ to the related keyframe model $\mathcal{S}^*$, the procedure is done incrementally in two stages.

The concept is as follows: primarily, given a reference sphere $\mathcal{S}^*$ and a candidate sphere $\mathcal{S}$, the cost function in (4.58) is employed to extract $\mathbf{T} = \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})$ and the entropy criteria

is applied for a similarity measure between the tuple $\{\mathcal{S}^*, \mathcal{S}\}$. While this metric is below a predefined threshold, the keyframe model is refined in a second stage – warping $\mathcal{S}$ and carrying out a geometric and photometric fusion procedure are composed of three steps:

- warping $\mathcal{S}$ and its resulting model error propagation
- data fusion with occlusions and outlier rejection
- an improved 3D point selection technique based on stable salient points

which are detailed in the following subsections.

5.4.2 Homogeneous Vector Uncertainty

Given a multivariate random variable $\mathbf{x}$ with mean μ_x and covariance Σ_x, such that $\mathbf{x} \sim D_1(\mu_x, \Sigma_x)$. For a mapping $\mathbf{y} = \mathbf{f}(\mathbf{x})$, it is possible to approximate the first two moments of $\mathbf{y}$ by just considering the first order approximation of $\mathbf{f}$ around $\mathbf{x}$ by taking the first two terms of the Taylor series expansion of $\mathbf{f}$ evaluated at μ_x and applying the expectation operator as follows:

$$\mathbf{y} \sim D_2(\mu_y, \Sigma_y), \quad \text{with} \quad \mu_y = \mathbf{f}(\mu_x) \text{ and } \Sigma_y = \mathbf{J}(\mu_x)\Sigma_x\mathbf{J}(\mu_x)^\top \tag{5.7}$$

In the general case with $\mathbf{z} = \mathbf{f}(\mathbf{x}, \ldots, \mathbf{y})$, assuming that $\mathbf{x}, \ldots, \mathbf{y}$ are independent:

$$\mu_z = \mathbf{f}(\mu_x, \ldots, \mu_y) \text{ and}$$

$$\Sigma_z = \mathbf{J}_x(\mu_x, \ldots, \mu_y)\Sigma_x\mathbf{J}_x(\mu_x, \ldots, \mu_y)^\top + \ldots + \mathbf{J}_y(\mu_x, \ldots, \mu_y)\Sigma_y\mathbf{J}_y(\mu_x, \ldots, \mu_y)^\top$$

This propagation holds exactly when $\mathbf{f}$ is linear for any distribution with bounded first two moments. To apply this parametrization, $\mathbf{f}$ must be smooth, with Σ being positive definite, *i.e.* $|\Sigma| > 0$.

5.4.3 Warped Sphere Uncertainty

An augmented spherical image $\mathcal{S} = \{\mathcal{I}, \mathcal{D}\}$ is composed of $\mathcal{I} \in [0, 1]^{m \times n}$ as pixel intensities and $\mathcal{D} \in \mathbb{R}^{m \times n}$ as the depth information for each pixel in $\mathcal{I}$. The basic environment representation consists of a set of spheres acquired over time together with a set of rigid transforms $\mathbf{T} \in \mathbb{SE}(3)$ connecting adjacent spheres (e.g. $\mathbf{T}_{ij}$ lies $\mathcal{S}_j$ and $\mathcal{S}_i$) – this representation is well described in [Meilland *et al.* 2011a].

The spherical images are encoded in a 2D image and the mapping between the image pixel coordinates $\mathbf{p}$ and depth to cartesian coordinates is given by $g : (u, v, 1) \mapsto \mathbf{q}$, $g(\mathbf{p}) = \rho\mathbf{q}_S$, with $\mathbf{q}_S$ being the point representation in the unit spherical space $\mathbb{S}^2$ and $\rho = \mathcal{D}(\mathbf{p})$ is radial depth. The inverse transform g^{-1} corresponds to the spherical projection model.

Point correspondences between spheres are given by the warping function w, under observability conditions at different viewpoints. Given a pixel coordinate $\mathbf{p}^*$, its coordinate $\mathbf{p}$ in another sphere related by a rigid transform $\mathbf{T}$ is given by a 3D screw transform, $\mathbf{q} = g(\mathbf{p}^*)$, followed by a spherical projection:

$$\mathbf{p} = w(\mathbf{p}^*, \mathbf{T}) = g^{-1}\left([\mathbf{I}\,\mathbf{0}]\mathbf{T}^{-1}\begin{bmatrix} g(\mathbf{p}^*) \\ 1 \end{bmatrix}\right), \tag{5.8}$$

where $\mathbf{I}$ is a (3×3) identity matrix and $\mathbf{0}$ is a (3×1) zero vector.

Warping the augmented sphere $\mathcal{S}$ generates a synthetic view of the scene $\mathcal{S}_w = \{\mathcal{I}_w, \mathcal{D}_w\}$, as it would appear from a new viewpoint. This section aims to represent the confidence of the elements in $\mathcal{S}_w$, which clearly depends on the combination of *an apriori* pixel position, the depth and the pose errors over a set of geometric and projective operations – the warping function as in (4.4). Starting with $\mathcal{D}_w$, the projected depth image is:

$$\begin{gathered} \mathcal{D}_w(\mathbf{p}^*) = \mathcal{D}_t(w(\mathbf{p}^*, \mathbf{T})) \text{ and } \mathcal{D}_t(\mathbf{p}) = \sqrt{\mathbf{q}_w(\mathbf{p}, \mathbf{T})^\top \mathbf{q}_w(\mathbf{p}, \mathbf{T})} \\ \text{with } \mathbf{q}_w(\mathbf{p}, \mathbf{T}) = \left([\mathbf{I}\,\mathbf{0}]\mathbf{T}\begin{bmatrix} g(\mathbf{p}) \\ 1 \end{bmatrix}\right) \end{gathered} \tag{5.9}$$

The uncertainty of the final warped depth $\sigma^2_{\mathcal{D}_w}$ then depends on two terms Σ_w and $\sigma^2_{\mathcal{D}_t}$; the former relates to the error due to the warping w of pixel correspondences between two spheres and the latter, to the depth image representation in the reference coordinate system $\sigma^2_{\mathcal{D}_t}$.

Before introducing these two terms, let's represent the uncertainty due to the combination of pose $\mathbf{T}$ and a cartesian 3D point $\mathbf{q}$ errors. Taking a first order approximation of $\mathbf{q} = g(\mathbf{p}) = \rho\mathbf{q}_S$, following section 5.4.3, the error can be decomposed as:

$$\Sigma_q(\mathbf{p}) = \sigma^2_\rho \mathbf{q}_S \mathbf{q}_S^\top + \rho^2 \Sigma_{q_S} = \frac{\sigma^2_\rho}{\rho^2} g(\mathbf{p}) g(\mathbf{p})^\top + \rho^2 \Sigma_{g(\mathbf{p})/\rho} \tag{5.10}$$

Depth information is usually extracted from a disparity map by a triangulation procedure as invoked in section (3.4.2.3), or directly retrieved by an active sensor. Dense stereo matching is quite a common technique to extract disparity maps. Operations during this extraction phase usually inherit random errors due to photometric information retrieved by the sensor itself (electronic noise) as well as systematic errors pertaining to the calibration phase. Adding up to that comes the problem of the disparity algorithm, which itself depends on other variables such as the cost function used, its related robustness or aliasing for example. Overhere, it is assumed that disparity follows: $d \sim \mathcal{N}(d, \sigma^2_d)$. The basic error model for the raw depth is given as: $\sigma^2_\rho \propto \rho^4$, which can be applied to both stereopsis and active depth measurement systems (for details more, the reader is referred to [Khoshelham & Elberink 2012]).

The next step consists of combining the uncertain rigid transform $\mathbf{T}$ with the errors

in $\mathbf{q}$. Given the mean of the 6DOF $\bar{\mathbf{x}} = \{t_x, t_y, t_z, \theta, \phi, \psi\}$ in 3D+YPR form and its covariance Σ_x, for $\mathbf{q}_w(\mathbf{p}, \mathbf{T}) = \mathbf{R}\mathbf{q} + \mathbf{t} = \mathbf{R}g(\mathbf{p}) + \mathbf{t}$,

$$\begin{aligned} \Sigma_{q_w}(\mathbf{p}, \mathbf{T}) = & \ \mathbf{J}_{\mathbf{q}}(\mathbf{q}, \bar{\mathbf{x}})\boldsymbol{\Sigma}_{\mathbf{q}}\mathbf{J}_{\mathbf{q}}(\mathbf{q}, \bar{\mathbf{x}})^\top + \mathbf{J}_{\mathbf{T}}(\mathbf{q}, \bar{\mathbf{x}})\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{J}_{\mathbf{T}}(\mathbf{q}, \bar{\mathbf{x}})^\top \\ = & \ \mathbf{R}\Sigma_q\mathbf{R}^\top + \mathbf{M}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{M}^\top, \end{aligned} \tag{5.11}$$

where Σ_q as in (5.10) and $\mathbf{M} \approx \begin{bmatrix} -y & z & 0 \\ \mathbf{I} & x & 0 & -z \\ & 0 & -x & y \end{bmatrix}$ for small rotations (for the general formula of $\mathbf{M}$, the reader is referred to [Blanco 2010]).

The first term Σ_w using (5.11) and (5.8) is given by:

$$\Sigma_w(\mathbf{p}^*, \mathbf{T}) = \mathbf{J}_{g^{-1}}(\mathbf{q}_w(\mathbf{p}^*, \mathbf{T}^{-1}))\Sigma_{q_w}(\mathbf{p}^*, \mathbf{T}^{-1})\mathbf{J}_{g^{-1}}(\mathbf{q}_w(\mathbf{p}^*, \mathbf{T}^{-1}))^\top \tag{5.12}$$

and $\mathbf{J}_{g^{-1}}$ is the jacobian of the spherical projection (the inverse of g). The second term expression for the depth represented in the coordinate system of the reference sphere using the warped 3D point in (5.11) and (5.9) is straightforward

$$\sigma^2_{\mathcal{D}_t}(\mathbf{p}^*, \mathbf{T}) = \mathbf{J}_{\mathcal{D}_t}(\mathbf{q}_w(\mathbf{p}^*, \mathbf{T}))\Sigma_{q_w}(\mathbf{p}^*, \mathbf{T})\mathbf{J}_{\mathcal{D}_t}(\mathbf{q}_w(\mathbf{p}^*, \mathbf{T}))^\top \tag{5.13}$$

with $J_{\mathcal{D}_t}(\mathbf{z}) = (\mathbf{z}^\top/\sqrt{\mathbf{z}^\top\mathbf{z}})$.

The uncertainty index $\sigma^2_{\mathcal{D}_w}$ is then the normalized covariance given by:

$$\sigma^2_{\mathcal{D}_w}(\mathbf{p}) = \sigma^2_{\mathcal{D}_t}(\mathbf{p})/(\mathbf{q}_w(\mathbf{p}, \mathbf{T})^\top\mathbf{q}_w(\mathbf{p}, \mathbf{T}))^2 \tag{5.14}$$

Finally, under the assumption of Lambertian surfaces, the photometric component is simply $\mathcal{I}_w(\mathbf{p}^*) = \mathcal{I}(w(\mathbf{p}^*, \mathbf{T}))$ and it's uncertainty $\sigma^2_{\mathcal{I}}$ is set by a robust weighting function on the error using a Huber's M-estimator as in [Meilland *et al.* 2011a].

5.4.3.1 Indoor spherical sensor model

The methodology presented above is generic to a spherical sensor. However, for the case of our spherical indoor sensor which is composed of Asus Xtion sensors, the depth information needs some additional post treatment because of uncertainties related to boundaries. It's principle of operation, which falls in the same category as Kinect-style sensors employs and infrared laser emitter for depth measurement but still makes use of the error prone disparity map as discussed earlier. An experimental set up was devised in [Nguyen *et al.* 2012] in order to model axial and lateral noises coming from the sensor. As expected, axial noise increases quadratically with the z-distance while lateral noise does not vary significantly with distance. However, the axial noise model holds good for a range of $10 - 60^0$ and beyond that, the noise model follow a rather hyperpolic function. In another investigation of [Khoshelham & Elberink 2012], the proportional tuning parameter of σ^2_ρ was found to be $\sigma^2_\rho = 2.05 \times 10^{-6}\rho^4$ (with $\rho \in [0, 5]$m) with 7cm error at an estimated confidence of 95.4% was obtained on the maximum range.

Besides model errors originating from disparity calculation, issues such as multiple refraction effects present in these active sensors, must also be taken into account. Areas

around boundary edges in the scene , where abrupt changes in the depth translates into spikes in the measurement. This problem is addressed in [Dryanovski *et al.* 2013], with a gaussian mixture model (GMM) of the relative depth over a neighbourhood for each pixel. Considering all the depths Z_{ij} $(i, j \in \mathbb{N})$ on a local window around $\mathbf{p} = [x\ y\ 1]^T$ such as $i \in [x-1, x+1]$, $j \in [y-1, y+1]$ and a gaussian kernel $\mathbf{W} = \frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$, the depth uncertainty considering the neighbourhood depth is then:

$$\sigma_{\rho_g}^2 = \sum_{i,j} w_{ij}(\rho_{ij}^2 + \sigma_{\rho_{ij}}^2) - \rho_g^2 \tag{5.15}$$

with $\rho_g = \sum_{i,j} w_{ij}\rho_{ij}$ and $\sigma_{\rho_{ij}}^2$ being the basic model. The model (5.15), as presented in [Dryanovski *et al.* 2013], improves the basic representation since it tackles the multiple refraction issues of Kinect-style IR sensors.

5.4.4 Probabilistic data association

The probabilistic data association method introduced overhere makes the use of the RGBD framework discussed in chapter (4) as well as the error model formulated in the previous section 5.4.3. An augmented spherical image $\boldsymbol{\mathcal{S}} = \{\boldsymbol{\mathcal{I}}, \boldsymbol{\mathcal{D}}\}$ is composed of $\boldsymbol{\mathcal{I}} \in [0, 1]^{m \times n}$ as pixel intensities and $\boldsymbol{\mathcal{D}} \in \mathbb{R}^{m \times n}$ as depth information for each pixel. In the proposed methodology, we start from a basic fact that a 3D point generated by the sensor is assumed to be coming from some random landmark whose true location is unknown. When the robot is first deployed, it needs to be familiarised with its immediate surrounding. Thus, the very first set of measurements obtained from the first frame are taken to be temporary landmark/features. Thereafter, observations coming from subsequent frames are compared to those initialised landmarks. Based on some decision boundaries, a match is then established between the measurement and observation values. If the two are in accordance with each other, the landmark's location is updated using the point's location. Temporary landmarks that match points consistently over many frames until the next keyframe initialisation are made permanent. With the help of the saliency map, the label of consistent and pertinent feature/landmark is added. The number of features/landmarks is limited to the size of our augmented sphere $\boldsymbol{\mathcal{S}}$.

5.4.5 Features/landmarks visibility scenario

Three types of visibility relationships are considered between the hypothesized depth map of the reference view and that of current views observed on the fly. Figure 5.10 illustrates those relationships when a reference $\boldsymbol{\mathcal{S}}^*$ is warped onto a current sphere $\boldsymbol{\mathcal{S}}$ and vice-versa. The point $\boldsymbol{\mathcal{P}}_k$ observed from $\boldsymbol{\mathcal{S}}^*$ is not observable in viewpoint $\boldsymbol{\mathcal{S}}$ as it is occluded by $\boldsymbol{\mathcal{P}}_m$ and hence considered as an outlier. On the other hand, the projection of $\boldsymbol{\mathcal{P}}_l$ from $\boldsymbol{\mathcal{S}}$ to $\boldsymbol{\mathcal{S}}^*$ results in $\boldsymbol{\mathcal{P}}_l$ being infront of $\boldsymbol{\mathcal{P}}_k$. There is a conflict between the measurement and the hypothesized depth since the emanating ray from viewpoint $\boldsymbol{\mathcal{S}}$ leading to $\boldsymbol{\mathcal{P}}_l$ violates the free space of $\boldsymbol{\mathcal{P}}_k$. $\boldsymbol{\mathcal{S}}^*$ would not have observed $\boldsymbol{\mathcal{P}}_k$ had there been a surface at $\boldsymbol{\mathcal{P}}_l$. This

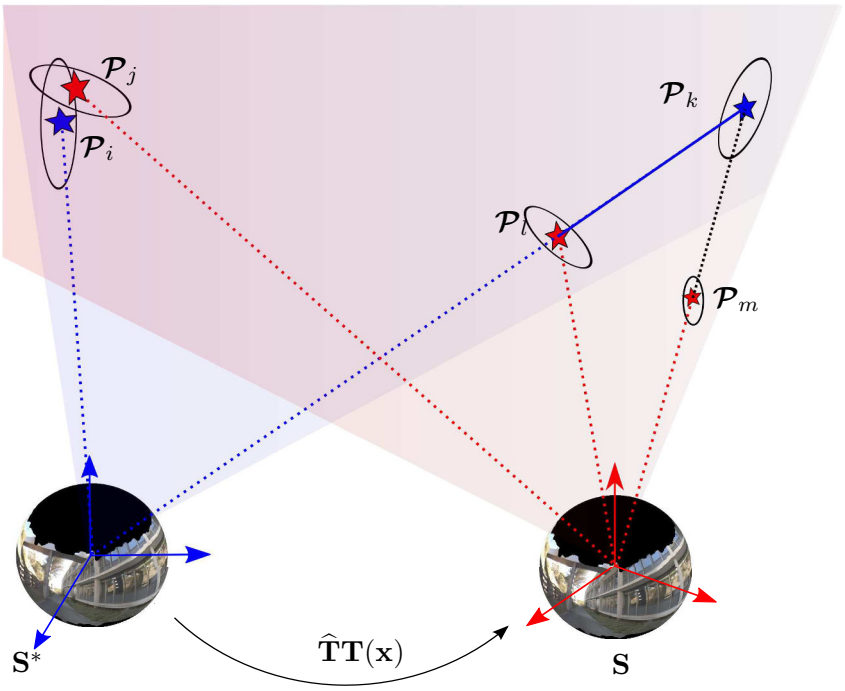

Figure 5.10: *probabilistic data association*

scenario is referred to as free space violation. Free space is the space between the origin of the sensor and its relative measurement. To conclude, we depict the case of an ideal scenario with the observation $\mathcal{P}_j$ considered as an inlier with respect to $\mathcal{P}_i$. In the following subsection, a method is presented to test each of the above estimates in order to select the most likely candidate for an inlier observation given a particular noisy measurement by considering all of the above spatial constraints. Eventually, the most consistent estimates are combined with their associated measurements to improve the measurement values so as to increase their likelihood of being seen accross the trajectory.

5.4.6 Formulation

Before combining the keyframe reference model $\mathcal{S}^*$ with that of the transformed observation $\mathcal{S}_w$, a probabilistic test is performed to exclude outlier pixel measurements from $\mathcal{S}_w$, allowing fusion to occur only if the raw observation agrees with its corresponding value in $\mathcal{S}^*$.

Hence, the tuple $\mathrm{A} = \{\mathcal{D}^*, \mathcal{D}_w\}$ and $\mathrm{B} = \{\mathcal{I}^*, \mathcal{I}_w\}$ are defined as the sets of model predicted and measured depth and intensity values respectively. Finally, let a class $c : \mathcal{D}^*(\mathbf{p}) = \mathcal{D}_w(\mathbf{p})$ relate to the case where the measurement value agrees with its corresponding observation value. Inspired by the work of [Murarka *et al.* 2006], the Bayesian framework for data association leads us to:

$$p(c|\text{A, B}) = \frac{p(\text{A,B}|c)p(c)}{p(\text{A,B})} \tag{5.16}$$

Applying independence rule between depth and visual properties and assuming a uniform prior on the class c (can also be learned from supervised techniques), the above expression simplifies to:

$$\begin{aligned} p(c|\text{A, B}) &\propto p(\text{A, B}|c)p(\text{B}|c) \\ &\propto p(\text{A}|c)p(\text{B}|c)p(c) \end{aligned}$$

$$\Rightarrow p(c|\text{A, B}) \propto p(\text{A}|c)p(\text{B}|c) \tag{5.17}$$

Treating each term independently, the first term of equation (5.17) is devised as $p(\text{A}|c) = p(\mathcal{D}_w(\mathbf{p})|\mathcal{D}^*(\mathbf{p}), c)$, whereby marginalizing over the true depth value ρ_i leads to:

$$p(\mathcal{D}_w(\mathbf{p})|\mathcal{D}^*(\mathbf{p}), c) = \int p(\mathcal{D}_w(\mathbf{p})|\rho_i, \mathcal{D}^*(\mathbf{p}), c)p(\rho_i|\mathcal{D}^*(\mathbf{p}), c)d\rho_i \tag{5.18}$$

Approximating both probability density functions as Gaussians, the above integral, following [Duda *et al.* 2001], reduces to:

$$p(\text{A}|c) \propto \exp\frac{-1/2(\mathcal{D}_w(\mathbf{p}) - \mathcal{D}^*(\mathbf{p}))^2}{\sigma^2_{\mathcal{D}_w}(\mathbf{p}) + \sigma^2_{\mathcal{D}^*}(\mathbf{p})} \tag{5.19}$$

Following a similar treatment,

$$p(\text{B}|c) \propto \exp\frac{-1/2(\mathcal{I}_w(\mathbf{p}) - \mathcal{I}^*(\mathbf{p}))^2}{\sigma^2_{I_w}(\mathbf{p}) + \sigma^2_{I^*}(\mathbf{p})} \tag{5.20}$$

Since equation (5.17) can be viewed as a likelihood function, it is easier to analytically work with its logarithm in order to extract a decision boundary. Plugging equations (5.19), (5.20) into (5.17) and taking its negative log gives the following decision rule for an inlier observation value:

$$\frac{(\mathcal{D}_w(\mathbf{p}) - \mathcal{D}^*(\mathbf{p}))^2}{\sigma^2_{\mathcal{D}_w}(\mathbf{p}) + \sigma^2_{\mathcal{D}^*}(\mathbf{p})} + \frac{(\mathcal{I}_w(\mathbf{p}) - \mathcal{I}^*(\mathbf{p}))^2}{\sigma^2_{I_w}(\mathbf{p}) + \sigma^2_{I^*}(\mathbf{p})} < \lambda^2_M, \tag{5.21}$$

relating to the square of the Mahalanobis distance. The threshold λ^2_M is obtained by looking up the χ^2_2 table.

Ultimately, we close up with a classic fusion stage, whereby depth and appearance based consistencies are coalesced to obtain an improved estimate of the spherical keyframe. Warped values that pass the test in (5.21) are fused up by combining their respective un-

certainties as follows:

$$\begin{aligned}\mathcal{I}^*_{k+1}(\mathbf{p}) &= \frac{\mathbf{W}^I_k(\mathbf{p})\mathcal{I}^*_k(\mathbf{p}) + \Pi_I(\mathbf{p})\mathcal{I}_w(\mathbf{p})}{\mathbf{W}^I_k(\mathbf{p}) + \Pi_I(\mathbf{p})}, \\ \mathcal{D}^*_{k+1}(\mathbf{p}) &= \frac{\mathbf{W}^D_k(\mathbf{p})\mathcal{D}^*_k(\mathbf{p}) + \Pi_D(\mathbf{p})\mathcal{D}_w(\mathbf{p})}{\mathbf{W}^D_k(\mathbf{p}) + \Pi_D(\mathbf{p})}\end{aligned} \tag{5.22}$$

for the intensity and depth values respectively and weight update:

$$\mathbf{W}^I_{k+1} = \mathbf{W}^I_k + \Pi_I \text{ and } \mathbf{W}^D_{k+1} = \mathbf{W}^D_k + \Pi_D \tag{5.23}$$

where $\Pi_I(\mathbf{p}) = 1/\sigma^2_{\mathcal{I}_w}(\mathbf{p})$ and $\Pi_D(\mathbf{p}) = 1/\sigma^2_{\mathcal{D}_w}(\mathbf{p})$ from the uncertainty propagation model of section 5.4.3.

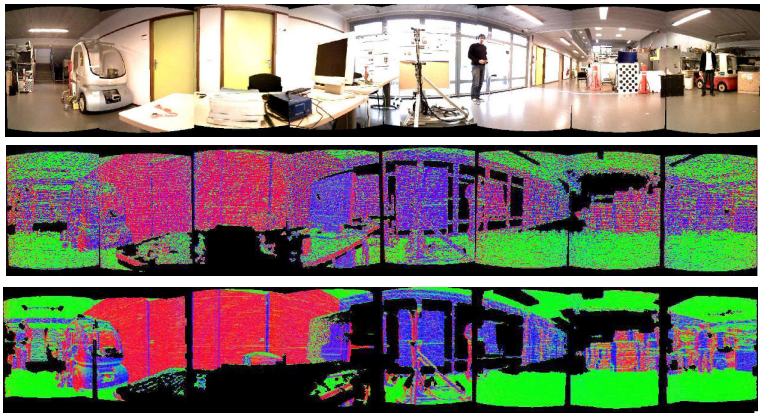

Figure 5.11: *Node comparison pre and post filtering*

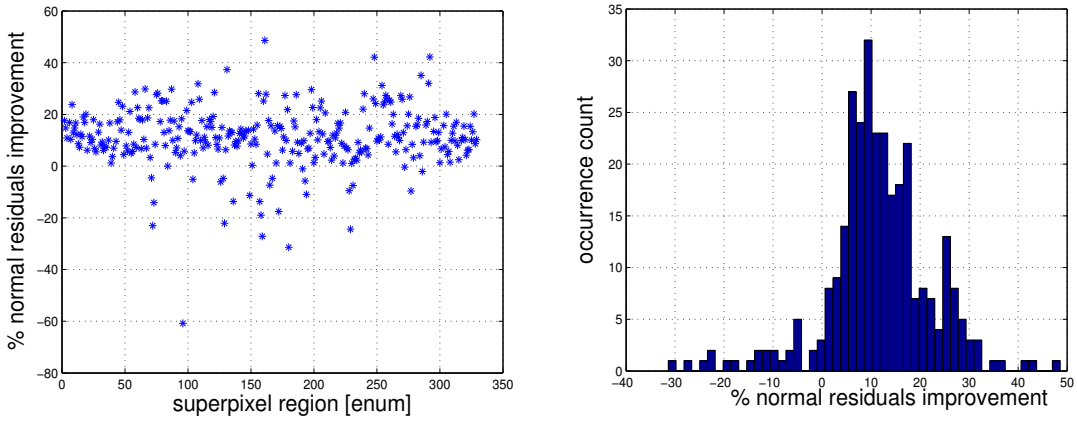

Figure 5.12: *Filtered depth map evaluation*

Figure 5.11 shows the normal consistency between raw and improved depth map of a spherical keyframe model of one of the nodes of the pose graph. The colours in the figure

Figure 5.13: *Sphere segmentation using SLIC superpixel algorithm*

encode surface normal orientations with patches belonging to the same surface exhibiting the same colour. In order to make a more qualitative evaluation in terms of the improvement achieved, a set of stages is defined where a metric quantity is extracted in order to justify the quality of the resulting depth map. For the sake of completeness, approached methodology is described overhere, corresponding partially to our work of depth map fusion while a more detailed synthesis is available in [Martins *et al.* 2015].

Spheres acquired in a window of n views are rasterised in a central reference frame and fused according to the above-mentioned technique described in this section. The fused and the raw depth map are segmented using the *simple linear iterative clustering* (SLIC) superpixel algorithm [Achanta *et al.* 2012], as shown in figure 5.13. Afterwards, for each segmented region $\mathcal{C}_i$ and an extracted planar model $\mathcal{M}_i$ for a set of points $\mathbf{q} \in \mathcal{C}_i$, the mean of the patch normals $\overline{\mathbf{n}}$ are then computed. Eventually, the following error metric is defined:

$$\mathcal{L}_n(\mathcal{C}_i, \mathcal{M}_i) = \int \int_{\mathbf{q} \in \mathcal{C}_i} \|\overline{\mathbf{n}} \bullet \mathbf{n}(\mathbf{q})\|_1 dq, \tag{5.24}$$

which basically gives a score $s_{\mathcal{L}_n}$ by computing the dot product of $\overline{\mathbf{n}}$ with all the normals $\overline{\mathbf{n}}_i$ of $\mathcal{C}_i$. Consequently, this results in a patch segmented depth map with each region now attributed a particular score. The same process is repeated for the original depth map, with $s^*_{\mathcal{L}_n}, \forall n \in \mathcal{M}$ and the percentage improvement between the two depth maps is obtained using the following ratio: $\frac{s_{\mathcal{L}_n} - s^*_{\mathcal{L}_n}}{s^*_{\mathcal{L}_n}}$. Figure 5.12(left) shows the annotated superpixel regions and their corresponding improvement achieved. Figure 5.12(right) highlights the same metric but better interpreted in a histogram representation. It is observed that the average improvement achieved is around 10% which goes up to 30% for certain regions.

5.4.7 Dynamic points filtering

So far, the problem of data fusion of consistent estimates in a local model has been addressed. But to improve the performance of any model, another important aspect of any

mapping system is to limit if not completely eliminate the negative effects of dynamic points. These points exhibit erratic behaviours along the trajectory and as a matter of fact, they are highly unstable. There are however different levels of "dynamicity" as mentioned in [Konolige & Bowman 2009]. Points/ landmarks observed can exhibit a gradual degradation over time, while others may undergo a sudden brutal change – the case of an occlusion for example. The latter being considerably apparent in indoor environments where small viewpoint changes can trigger a large part of a scene to be occluded. Other cases are observations undergoing cyclic dynamics (doors opening and closing). Whilst the above-mentioned behaviours are learned in clusters [Konolige & Bowman 2009], in this work, points with periodic dynamics are simply evaluated as occlusion phenomena.

Besides dynamic points pertaining to object boundaries embedded in the general scene, there is a subcategory of points –those around the scene's border areas, too contribute to undesirable effects. A boundary point in a certain viewpoint $\mathcal{V}_i$ can be either warped inside another viewpoint $\mathcal{V}_j$ resulting in either a non boundary entity or simply outside $\mathcal{V}_j$. Behaviours discussed above, though partly handled by robust Visual Odometry cost functions (section 4.4), do however contribute to biased motion estimates [Zhao *et al.* 2005]. In the following part, a description of the mechanism behind dynamic points filtering is discussed.

The probabilistic framework for data association developed in the section 5.4.6 is a perfect fit to filter out inconsistent data. 3D points giving a positive response to test equation (5.21) are given a vote 1, or otherwise attributed a 0. This gives rise to a confidence map $\mathcal{C}_i^*(k)$ which is updated as follows:

$$\mathcal{C}_i^*(k+1) = \begin{cases} \mathcal{C}_i^*(k) + 1, & \text{if } \lambda_M^{(95\%)} < 5.991 \\ 0, & \text{otherwise} \end{cases} \tag{5.25}$$

Hence, the probability of occurence is given by:

$$p(\text{occur}) = \frac{\mathcal{C}_i^*(k+N)}{N}, \tag{5.26}$$

where N is the total number of accumulated views between two consecutive keyframes. $p(\text{occur})$, though it gives an indication on how many times a point has been tracked along the trajectory, it can however not distinguish between noisy data or an occlusion. Treading on a similar technique to that adopted in [Johns & Yang 2014], a Markov observation independence is imposed. In the event that a landmark/3D point has been detected at time instant k, it is most probable to appear again at $k+1$ irrespective of its past history. On the contrary, if it has not been re-observed, this may mean that the landmark is quite noisy/unstable or has been removed indeterminately from the environment and has little chance to appear again. These hypotheses are formally translated as follows:

$$\gamma_{k+1}(\mathbf{p}^*) = \begin{cases} 1, & \text{if } \mathbf{p}_k^* = 1 \\ (1 - p(\text{occur}))^n, & \text{otherwise} \end{cases} \tag{5.27}$$

Finally, the overall stability of the point is given as:

$$p(\text{stable}) = \gamma_{k+1} p(\text{occur}) \tag{5.28}$$

5.4.8 Application to Saliency map

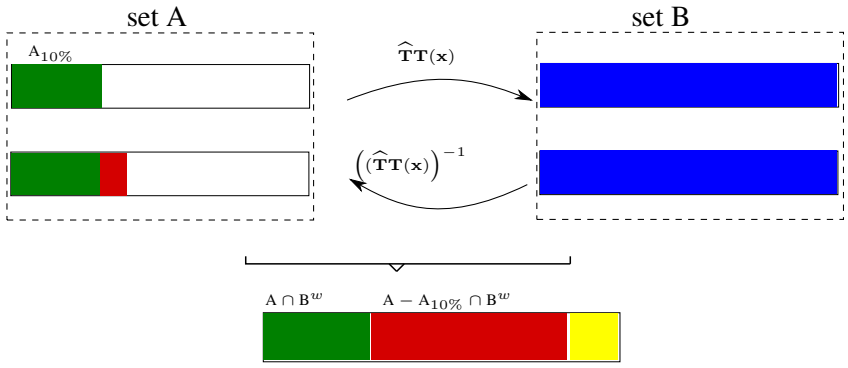

Figure 5.14: *Saliency map skimming*

Instead of naively dropping out points below a certain threshold, for e.g, $p(\text{stable}) < 0.8$, they are better pruned out of a saliency map 4.2.4. A saliency map, S^*_{sal}, is the outcome of careful selection of the most informative points, best representing a 6 degree of freedom pose, $\mathbf{x} \in \mathbb{R}^6$, based on a Normal Flow Constraint spherical jacobian. The underlying algorithm is outlined below: The green and red sub-block in figure (5.14) presents the set of inliers and outliers respectively, while the yellow one corresponds to the set of pixels which belong to the universal set $\{\mathcal{U} : \mathcal{U} = \text{A} \cup \text{B}^w\}$ but which have not been pruned out. This happens when the Keyframe criteria based on an entropy ratio α [Kerl *et al.* 2013b] is reached. The latter is an abstraction of uncertainty related to the pose $\mathbf{x}$ along the trajectory, whose behaviour shall be discussed in the results section.

The novelty of this approach compared to the initial work of [Meilland *et al.* 2010] is two-fold. Firstly, the notion of uncertainty is incorporated in spherical pixel tracking. Secondly, as new incoming frames are acquired, rasterised and fused, the information content of the initial model is enriched and hence the saliency map needs updating. This gives a newly ordered set of pixels to which is attributed a corresponding stability factor. Based on this information, an enhanced pixel selection is performed consisting of pixels with a greater chance of occurence in the subsequent frame. This set of pixel shall then be used for the forthcoming frame to keyframe motion estimation task. Eventually, between an updated model at time t_0 and the following re-initialised one, at t_n, an optimal mix of information sharing happens between the two.

Algorithme 3 3D Points Pruning using Saliency map

Require: $\{\mathcal{S}^*{}_{sal}, \mathcal{C}^*_i(k), N, n\}$
  return Optimal Set $\mathbf{A}_{10\%} \in \mathcal{S}^*{}_{sal}$
  Initialise new $\mathbf{A}$
  for i=$\mathcal{S}^*{}_{sal}(\mathbf{p}^*) = 1$ **to** $\mathcal{S}^*{}_{sal}(\mathbf{p}^*) = \max$ **do**
    compute $p_{(\text{occur})}(\mathbf{p}^*_i)$
    compute $\gamma_{k+1}(\mathbf{p}^*_i)$
    compute $p_{(\text{stable})}(\mathbf{p}^*_i)$
    if $p_{(\text{stable})}(\mathbf{p}^*_i) \geq 0.8$ **then**
      $\mathbf{A}[\,i\,] \longleftarrow \mathbf{p}^*_i$
    if length$(\mathbf{A}[\,i\,]) \geq \mathbf{A}_{10\%}$ **then**
      break

5.4.9 Results

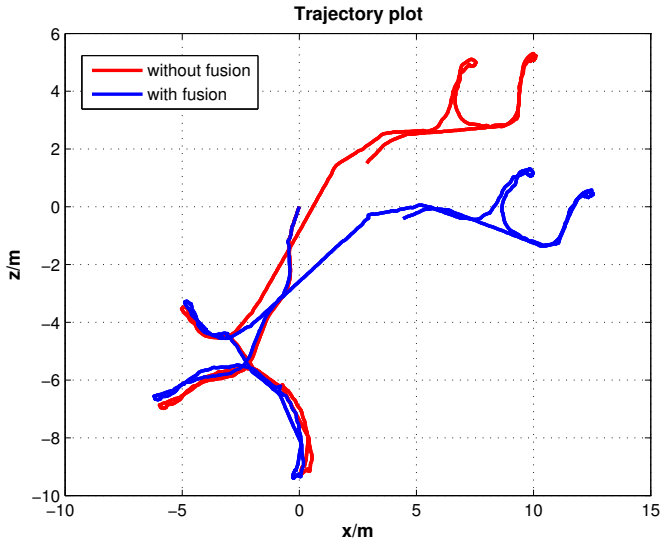

Figure 5.15: *Trajectory comparison with and wihout fusion using error model*

The same experimental set up is devised similar to section 5.3.1 to evaluate the contribution of the proposed methodology (*cf.* pipeling of figure 5.9). Figure 5.15 illustrates the trajectories obtained from two experimented methods, namely; RGBD registration without (***method 1***) and with keyframe fusion (***method 3***) in order to identify the added value of the fusion stage. The noticeable trajectory discrepancies between ***method 1*** and ***method 3*** suggest that erroneous pose estimations which have previously occured with the former technique have well been suppressed with the latter fusion method. This is even more emphasized by visually inspecting the 3D structure of the reconstructed environment as shown in figure (5.16) where, again, the two maps are compared side by side. This time reconstruction using the newly devised method contributes significantly to drift reduction. As it

Figure 5.16: *Reconstruction comparison with Kahn dataset*

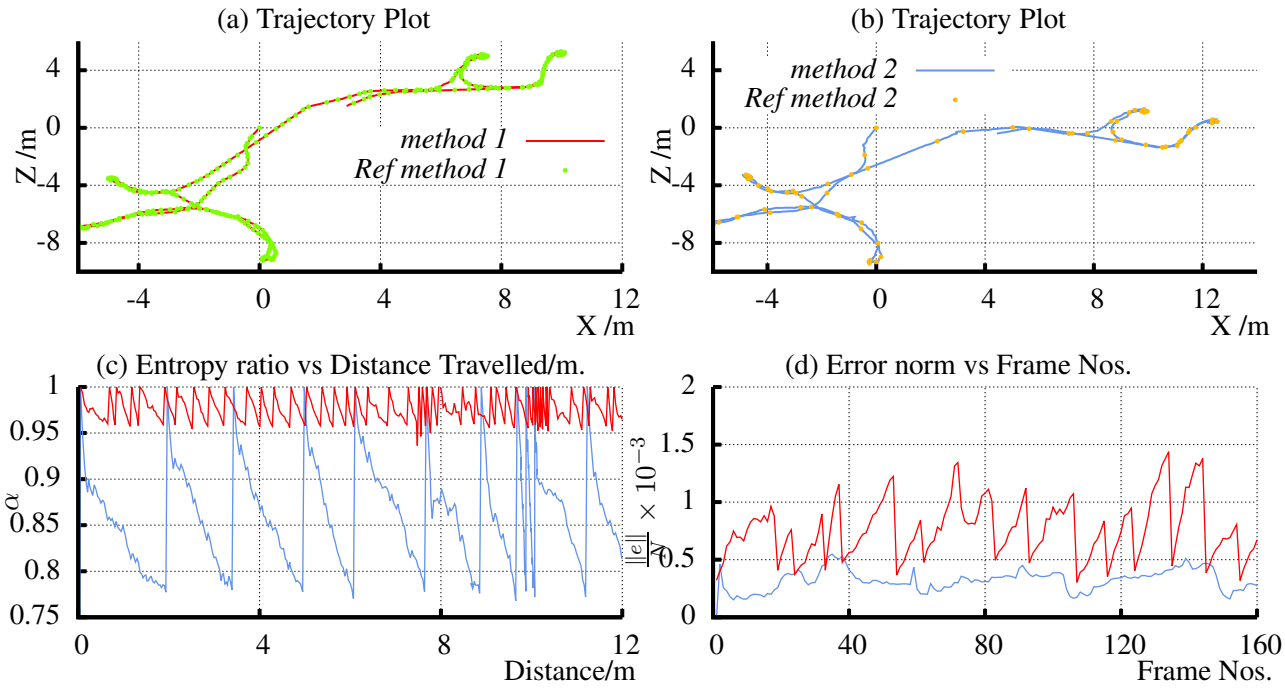

Figure 5.17: *Performance comparison with Kahn0 dataset*

is observed, the overall global reconstruction has well been enhanced by the suppression of outlying duplicated wall structures resulting in a much improved alignment. Out of the 270 keyframes initially recorded for ***method 1***, only **67** key spheres were retained for *method 3*, representing a net reduction of 75.2%. Here again, the fact that using less keyframes eventually does reduce accumulated errors resulting from pose compositions in the graph

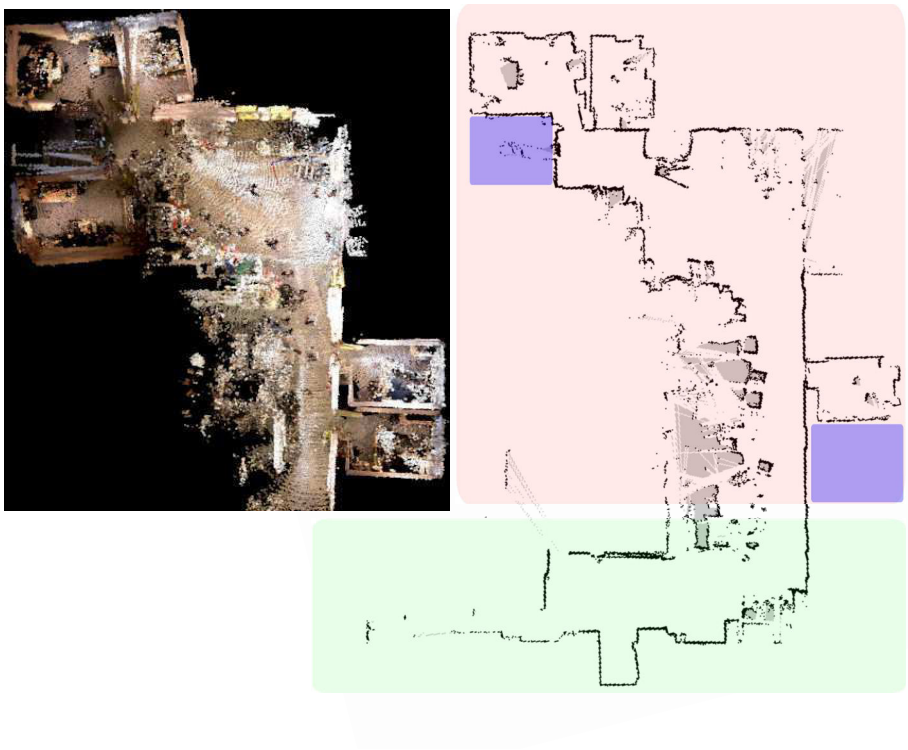

Figure 5.18: *Comparison between vision and laser maps*

building process.

Finally, figure 5.17(a), (b) illustrate the total trajectory travelled in the building with the spherical keyframes for *method 1* and *method 3* respectively. The gain in compactness for *method 3* is clearly demonstrated by the sparse positioning of keyframes in the skeletal pose graph structure. Figure 5.17(c) depicts the behaviour of our keyframe selection entropy-based criteria α whose threshold for α is heuristically tuned. For *method 1*, a preset value of 0.96 was used based on the number of iterations to convergence of the *photometric+geometric* cost function outlined in section 4.4. With the fusion stage, the value of α was allowed to drop to 0.78 with generally faster convergence achieved. Figure 5.17(d) confirms the motion estimation quality of *method 3* as it exhibits a lower error norm across frames as compared to *method 1*. Accordingly, this result justifies the improvement made with the depth map fusion process. Figure 5.18 shows a side by side comparison between RGBD map reconstruction with *method 3* and a map obtained using a 2D laser scanner. The Green shade represents region not visited with the RGB-D sensor whilst blue shades indicates regions not mapped with the laser sensor. As observed, the overall building structure is preserved using *method 3*. To conclude, table 5.1 summarises the performance obtained with *methods 1, 2 and 3*.

	Method 1	*Method 2*	*Method 3*
Keyframe criteria α	0.96	0.96	0.78
Keyframe reduction (%)	–	6.7	**75.2**
Mean convergence error	0.5889	0.5081	**0.2413**
Mean nos. iterations	28.3	27	**23.5**

Table 5.1: *Methods comparison*

5.4.10 Discussion

Lower convergence rate of the cost function cannot always be guaranteed since it depends on various factors such as its high non-linearity (e.g warping) which does not always lead to a global minimum, occlusions phenomena in the scene, or even lack of relevant information from the sensor due to its range limitations. Door passing or natural light reflection from window panes for example can create biased motion estimates. Therefore, in cases of local minima, an upper convergence threshold is set and a keyframe is re-initialised with the previous acquired frame and its corresponding motion estimate from the model.

Trajectory drift has been reduced, but not completely eliminated. This is partly attributed to the error model worked out in this work. The error model assumed ideally perfect spherical RGB-D images and hence calibration errors were ignored. Furthermore, the optimal set of 3D point based on the saliency map is obtained between two respective spherical models, updated along the trajectory. It would however be interesting to study the behaviour of these sets along different timescales for example. Therefore, we believe that this work has definitely given a clear direction for improvements.

5.5 Conclusion

In this chapter, different techniques have been proposed to suppress problems arising due to measurement and observation errors. In order to compare measurements in space, they need to be transformed in single spatial representative frame. In our case, it was chosen to be the reference keyframe. Bringing back observations at different viewpoints exposes issues such as occlusions and disocclusions occurring from surfaces appearing and disappearing in between frames. A first approach was considered whereby a rupture model was considered on the discrete signal of the depth map. This model is based on a Page-Hinckley test which detects changes based on the signal's mean-time series variation. Experiments with this first approach showed that the problem of drift has been poorly tackled due to several reasons. Firstly, analysing only signal variations has certain disadvantages such as false and delayed detections. Moreover, the uncertainty of the observed depth map has not been considered taking into account the effect of warping. Finally, during the filtering phase, only the depth information has been considered in the weighting average filter ignoring the contribution of the photometric information. All these loopholes in the first

approach explains the inconsistency in the overall global map reconstruction. In order to reduce the effect of the above mentioned issues, a back-end SLAM solution has been introduced by optimising over poses only, using manual loop closures to understand the effect caused by local pose graph correction. This technique has shown to be vital in the implementation of a full SLAM system. However, due to implementation issues, coming from the incompatibility of the dissimilar development platforms (Matlab and C++), this idea was later shelved due to ongoing coding developments.

Nevertheless, identified caveats of the first approach was used as a springboard to propose an improved methodology. Measurements obtained from the sensor are inherently noisy and adding up to that, dynamic errors introduced by further manipulation of the depth map estimate, if not handled properly, incorporate erroneous measurements in the depth map which are thereafter propagated along the visual odometry chain leading to erroneous pose estimates. This is very much undesirable in our pose graph representation as only a wrong estimate may single handedly disrupt the whole graph. Therefore in a second approach, we considered the explicit propagation of the depth map by taking into account all the transformations that a particular observation is subjected to until it represented in the reference keyframe of interest. At the end of this process, for each observed depth map, its associated uncertainty map is also deduced. This component is vital in the implementation of our probabilistic framework for data association. This time, both photometric and geometric information together with their associated measurement uncertainties are considered. Eventually, data fusion leads to a depth map improvement of $10\% - 30\%$. It should be pointed out that systematic errors pertaining to sensor calibration has not been modelled. Furthermore, the aspect of dynamic points is treated with application to the saliency map. By reshuffling the saliency map with the additional notion of uncertainty, an improved point selection is achieved based on the stability concept introduced which predicts the behaviour of points in the reference frame. The overall performance is reflected in our estimation algorithm with comparatively lower estimation errors are achieved resulting to a more consistent map. At the end of this work our set of augmented spheres now consists of two more entities; the uncertainty and the stability maps.

CHAPTER 6

Conclusion and Perspectives

6.1 Conclusion

In this thesis, a vision only SLAM framework is presented, built around a spherical egocentric environment representation. Focus is made on a pose graph representation whereby nodes are connected with edges, established by spherical VO. The wide angle 360^0 system configuration provides two major benefits; an enriched model required for accurate localisation and second, the use of keyframes gives more compactness which is an important aspect for vast scale exploration. A direct dense based approach for registration is used, preferred to erroneous feature based techniques due to better achievable precision by using all the information content output by the sensor. This eliminates the requirement for an additional feature detection step prior to registration. Each node in our keyframe based representation consists of an augmented sphere made up three entities; a spherical RGB image, a spherical depth map and a saliency map. The advantage of the saliency map is that it provides and additional pixel information based on the Normal Flow Constraint leading to an arrangement of pixels which best condition a 6 DOF motion constraint. Therefore, instead of using the entire RGBD information for registration, only a subset of $5 - 10$ percent is injected in the optimisation cost function, thereby reducing computational cost. This new trend is termed as semi-direct registration in literature.

The first objective of our work was to improve frame to keyframe registration. Initially based on a dense based registration technique, this approach has certain limitations; poorly textured areas, considerable illumination fluctuation across wide viewpoints lead to feature mismatching, hence to erroneous pose estimate. In this context, a hybrid cost function has been proposed, which includes both photometric and geometric information in a single framework. Our second contribution was to improve on the keyframe selection criteria in the pose graph construction process. A previous criteria based on MAD was implemented in the system, which depended on the photometric residual of frame to keyframe registration. The MAD is highly sensible on illumination changes between frames. While the criteria may work well for small interframe displacements with minimal lighting variations, considerable photometric changes along a driven trajectory leads to redundant accumulation of frames. To tackle this problem, a second criteria has been implemented based on differential entropy. The latter is an abstraction of the pose uncertainty from the motion estimate. This new criteria works better than the MAD in reducing keyframe redundancy, resulting in reduced integration of tracking errors.

Our algorithm was tested on four types of dataset; one synthetic, two indoors and one outdoor. Results obtained with the Inria Kahn dataset (indoor) exposed some weaknesses of the algorithm which required further investigation. Though we do not have a ground truth comparison, the 3D point cloud reconstruction of the environment revealed inconsistencies in the local map. One of the main possibilities evoked is VO failures along the trajectory. Door passing is a critical issue with the indoor spherical sensor. Due to its range limitation regions with no observation in the depth map can lead to a bias pose estimate induced and propagated in the map. To be able to anticipate such discrepancies, a preliminary approach was devised based on metric loop closure.

Our algorithm was further tested on an outdoor urban environment. Due to the extremely noisy depth map computed from passive stereo vision techniques, spherical VO failure was imminent. Consequently, before applying spherical VO, the depth maps were subjected to a prefiltering phase. VO was run again for trajectory computation. This time, a far more consistent trajectory is obtained even though several failures were registered in the map. These occurred mainly when the vehicle negotiated curbs, registering significant changes between frames causing failures mainly attributed to rotation estimation. Nevertheless, the overall noticed trajectory was locally "piecewise" consistent. However, all the above-mentioned issues can be corrected using a proper SLAM back-end framework with optimisation over the graph and the structure.

In the second fold of our work, we focussed on ways to improve the sensor information. In an initial investigation, observations pertaining to various frames acquired along the trajectory were brought to the coordinate frame of their nearest neighbour keyframe. Discrete depth maps, stacked in a single representation results in a signal flow for each pixel, describing the profile of an observed surface across different viewpoints. In order to detect noise and occlusion phenomena, the Page-Hinckley test raises an alarm whenever a model rupture is detected. Consistent measurements were fused up to improve the depth map of the reference frame. Though drift has been reduced, no significant gain in the overall map reconstruction was noted. The identified loopholes in the first methodology was used to propose a more consolidated approach whereby sensor errors were properly modelled. The devised probabilistic data association framework together with the treatment of dynamic points led to a considerable improvement in the overall computed trajectory, hence the local map reconstruction. The idea of fusing photometric and geometric information taking into account motion and sensor uncertainty led to two major benefits. Primarily, better motion estimates were obtained and secondly, less keyframes were registered, representing a compactness of 75%. Finally, we have augmented our local node information with two additional entities; the uncertainty map (as applied to photometry and geometry) and the stability map.

6.2 Perspectives

For the front-end VSLAM framework presented in this work to be fully operational on a mobile robotic platform, several additional aspects needs to be considered. First and for most, parallely with pose graph construction, a back-end SLAM module needs to be integrated. Though we have conducted an initial investigation through metric loop closure, the ideal way would be to detect loop closures at a topological level, such as the bag of words technique of [Chapoulie *et al.* 2011] or using other appearance based techniques such as FABMAP [Paul & Newman 2010] or that of [Johns & Yang 2014]. Once loop closure is detected on purely appearance based level, this constraint can be enforced in using pose graph optimisation technique of [Kümmerle *et al.* 2011].

At the level of spherical odometry, there is still plenty room for improvements. In our hybrid formulation, an illumination model has not been considered. It would be interesting to model illumination changes in the cost function to provide more robustness to photometric changes across wide viewpoints. Furthermore, the two cost functions implemented in the hybrid optimisation framework has been tuned in a heuristic style, similar to [Henry *et al.* 2012]. At the time of writing this manuscript, two interesting techniques are under investigation. The first one is based on the modelling of the augmented cost function using a bivariate T-distribution in a way to that of [Kerl *et al.* 2013a]. The second one involves incorporating the uncertainty maps of both photometry and geometry into the cost function, in a similar fashion to [Engel *et al.* 2014].

Though we have partly tackled the problem of localisation whilst applying metric loop closure on the indoor dataset, the improved pose graph has not been explicitly tested. It would be interesting to evaluate the graph by taking an arbitrary frame in the dataset and try to localise with respect to its closest keyframe in the pose graph. However, this process is not straightforward and require a two stage initialisation process. This is due to the convergence properties of direct methods which requires a proper initial guess. Overhere, a feature based technique can be implemented by extracting salient points pertaining to the current frame as perceived by the robot and to its nearest identified keyframe (after and initial identification of apperance based correspondence using the above mentioned techniques). A rough pose estimate can be computed by using Horn's RANSAC [Horn 1987] method to provide and initial 3D transformation. The latter can then be refined using our direct approach.

Once an accurate localisation is obtained, the map can be updated using the proposed probabilistic framework. In this probabilistic framework, only photometric and depth information has been modelled. It would be interesting to consider the normal map propagation as well and including it accordingly, in a similar way to [Herbst *et al.* 2011].

As mentioned in the conclusion section, each node in our graph is now augmented with uncertainty and stability maps. This designed framework encompassing points' visibility across the visual trajectory is immediately adaptable to the long term and short term memory reasoning of [Dayoub *et al.* 2011] or its recent extension [Bacca 2012]. At each

keyframe node, the set of re-arranged points based on the saliency map can be viewed as an initialisation of the long term memory (LTM), with the short term memory (STM) originally declared empty. The STM and the LTM can then be refined over multi mapping sessions (through various acquisition campaigns at different timescales). Finally, when exploring over various timescales, the initial graph constructed may be subjected to changes. Nodes can be added to the graph when a new place in the the environment appears. To prevent graph to bulge out by the addition of redundant nodes, the proposition of [Kretzschmar *et al.* 2010] can be used to address pose graph pruning.

Conclusions et Perspectives

Conclusions

Dans cette thèse nous avons présenté un cadre pour du SLAM uniquement basé sur la vision construit sur une représentation sphérique égocentique de l'environnement. Une attention toute particulière a été accordée à la représentation par graphe de pose dans laquelle des nœuds connectés par des arêtes sont établis par l'odométrie visuelle sphérique. La configuration large angle à 360^0 du système avance deux bénéfices majeurs, à savoir un modèle enrichi nécessaire au positionnement précis, et la compacité augmentée produite par l'utilisation des images-clés, qui est alors un aspect important pour l'exploration à grande échelle. Une approche directe basée sur la densité est utilisée pour la détection, préférée aux techniques basées sur les caractéristiques erronées pour la raison de la possibilité d'une meilleure précision obtenue en utilisant toute l'information obtenue du capteur. Ceci élimine le besoin d'une étape de de détection de caractéristiques particulières supplémentaire et préliminaire à la détection elle-même. Chaque nœud de notre représentation à base d'images-clés correspond à une sphère visuelle augmentée, constituée des trois éléments que sont une image sphérique RGB, une carte sphérique de profondeur, et une carte d'intérêt. L'avantage de la carte d'intérêt est d'apporter une information pixellaire supplémentaire issue de la Contrainte de Flux Normal (Normal Flow Constraint), et conduit à un arrangement des pixels idéal pour le conditionnement d'une contrainte de déplacement à 6 degrés de libertés. Ainsi, plutôt que d'utiliser toute l'information RGBD pour la détection, seul un sous-ensemble de 5–10 % est utilisé dans l'optimisation de la fonction de coût, réduisant ainsi le coût calculatoire. Dans la littérature, cette nouvelle philosophie est appelée détection semi-directe.

Le premier objectif de notre travail a été d'améliorer la détection d'image à image-clé. Originant d'une technique de détection par densité, cette approche a certaines limites : en milieu peu texturé, ou à l'illumination aux importantes modifications sur de larges champs de vision, elle conduit à des caractéristiques particulières erronées, donc à une estimation de la pose fausse. Dans ce contexte, une fonction de coût hybride a été proposée, laquelle lie les informations photométriques et géométriques en une unique approche. Notre seconde contribution a été l'amélioration du critère de sélection des images-clés du processus de construction du graphe de pose. Un précédent critère appuyant sur la méthode MAD a été implémenté, il dépendait du résidu photométrique de la détection d'image à image-clé. La MAD est hautement sensible aux modifications de luminosité entre les images. Ainsi, alors que ce critère peut-être des plus pertinent pour de petits déplacements entre deux images successives et pour des variations de luminosité minimales, de grandes variations photométriques le long d'une trajectoire conduit à une accumulation redondante des scènes. Pour enrayer ce problème, une second critère utilisant l'entropie différentielle a été implémenté. Celle-ci est une abstraction de l'incertitude de la pose due à l'estimation du déplacement.

Ce nouveau critère est bien plus efficace que celui utilisant la MAD dans la réduction des redondances des images-clés, permettant ainsi l'obtention de l'intégration réduite des erreurs de dérive.

Notre algorithme a de plus été testé en environnement extérieur urbain. De part la carte de profondeur hautement bruitée calculée à partir de techniques de vision stéréo passives, l'échec de la VO sphérique était attendu. En conséquence de quoi, avant le traitement de la VO sphérique, un préfiltrage des cartes de profondeur a été appliqué. Après ce traitement, la VO est relancée et cette fois une trajectoire beaucoup plus cohérente est obtenue, malgré quelques échecs sur la carte. Ces échecs interviennent principalement lorsque le véhicule négociait des trajectoires courbes, impliquant des changements notables entre les images principalement causés par l'estimation de la rotation. En dépit de ces échecs, la trajectoire relevée était localement cohérente, « par morceaux ». Tous ces problèmes ici mentionnés peuvent être corrigés par l'utilisation de ce schéma dans un système de SLAM dédié avec optimisation sur le graphe et la structure.

Dans la seconde partie de notre travail, nous nous sommes concentrés sur l'amélioration de l'information obtenue des capteurs. Dans notre première analyse, différentes observations dans différents référentiels obtenus le long de la trajectoire étaient transformé jusqu'au référentiel de l'image-clé voisine la plus proche. Les cartes de profondeur discrètes résultant en un empilement de flux de signaux pour chaque pixel, décrivent le profil de la surface observée selon différents points de vues. Pour détecter le bruit et les occultations, le test de Page-Hinckley signale un problème dès que une cassure du modèle est détectée. Les mesures cohérentes ont été fusionnées pour améliorer la carte de profondeur du référentiel. Bien que la dérive aie été réduite, pas de gain significatif dans la reconstruction de la carte générale n'a été relevé. Les problèmes de la première méthode ont permis de proposer une approche plus robuste dans laquelle les erreurs de capteurs ont été correctement modélisées. La méthode d'association probabiliste de données imaginé, de pair avec le traitement des points dynamiques, ont permis une amélioration considérable de la trajectoire globale calculée, et donc de la reconstruction de la carte locale. L'idée de l'association des données photométriques et géométriques, prenant en compte les erreurs des déplacements et des capteurs, a permis deux principales avancées. D'une part, de meilleures estimations de déplacement sont obtenues, et d'autre part, moins d'images-clés sont accumulées, pour une compacité de l'ordre de 75%. Finalement, nous avons augmenté les informations associées aux nœuds locaux de deux entités supplémentaires, que sont les cartes d'incertitudes (telle que s'appliquant à la photométrie et à la géométrie), et de stabilité.

Perspectives

Pour que la méthode d'entrée du VSLAM présentée dans ce travail soit complètement opérationnelle sur une plate-forme robotique, d'autres aspects doivent être considérés. Le premier et plus important point avec la construction du graphe de pose, vient la méthode traitement SLAM qui doit être intégré. Bien que nous ayons mené une étude préliminaire

par la fermeture de boucle métrique, une meilleure approche serait de détecter les fermetures de boucles par une approche topologique, telle que la technique du sac de mots de [Chapoulie *et al.* 2011], ou alors en utilisant d'autres techniques basées sur la forme telles que FABMAP [Paul & Newman 2010] ou encore celle de [Johns & Yang 2014]. Une fois qu'une fermeture de boucle est détectée du point de vue purement de la forme, la contrainte peut alors être appliquée par l'utilisation de la technique d'optimisation du graphe de pose de [Kümmerle *et al.* 2011].

Au niveau de l'odométrie sphérique, de nombreuses améliorations sont possibles. Dans notre formulation hybride, nous n'avons pas considéré de modèle d'illumination. Il serait donc opportun de modéliser les changements d'illumination dans la fonction de coût pour donner plus de robustesse aux modifications photométriques au travers des différents points de vue. De plus, les deux fonctions de coût implémentées dans le module d'optimisation hybride ont été ajustées de manière heuristique, de manière similaire à [Henry *et al.* 2012]. Au moment de l'écriture de ce manuscrit, deux techniques prometteuses sont étudiées. Une première est basée sur la modélisation de la fonction de coût augmentée par l'utilisation d'une distribution en T bivariée telle que présenté par [Kerl *et al.* 2013a]. La seconde intègre les cartes d'incertitudes autant de la photométrie que de la géométrie dans la fonction de coût, comme présenté cette fois par [Engel *et al.* 2014].

Bien que nous ayons partiellement résolu le problème de la localisation tout en appliquant une fermeture de boucle métrique sur l'ensemble de données en intérieur, le graphe de pose amélioré n'a pas été testé exhaustivement. Il serait intéressant de tester ce graphe en prenant une image arbitraire de l'ensemble de données et d'essayer de se localiser vis-à-vis de l'image-clé la pus proche du graphe de pose. Toutefois cette démarche n'est pas des plus simples et nécessite une procédure d'initialisation en deux étapes. Ceci est dû aux propriétés de convergence des méthodes directes qui requièrent une estimation initiale sensée. Sur le même sujet, une technique basée sur les caractéristiques particulières peut être implémentée par l'extraction des points d'intérêt propres tant à l'image courante telle que perçue par le robot, que son image-clé identifiée la plus proche (après l'identification initiale d'une correspondance basée sur l'apparence en utilisant les techniques mentionnées plus haut). Une estimation grossière de la pose peut être calculée par l'utilisation de la méthode RANSAC de Horn [Horn 1987], produisant une première transformation 3D. Cette transformation peut ensuite être affinée en utilisant notre approche directe.

Une fois qu'un positionnement précis est obtenu, la carte peut être mise à jour en utilisant la méthode probabiliste proposée. Dans cette méthode, seules les distributions des données photométriques et de profondeur ont été modélisées. Il serait intéressant de considérer aussi la propagation normale à la carte, pour l'inclure en conséquence comme exemplifier par [Herbst *et al.* 2011].

Comme mentionné dans la conclusion, précédente section, chaque nœud de notre graphe est maintenant accompagné des cartes d'incertitude et de stabilité. Cette méthode, contenant la visibilité des points le long de la trajectoire visuelle, est immédiatement adaptable au raisonnement de mémorisations à long et court termes de [Dayoub *et al.* 2011],

ou encore à celui de sa récente extension [Bacca 2012]. Dans ce contexte, à chaque nœud image-clé tous les points réorganisés en référence à la carte d'intérêt peuvent être vus comme des initialisation de la mémoire à long terme (long term memory, LTM), pendant que la mémoire à court terme (short term memory, STM) est initialisée vide. Les STM et LTM peuvent ensuite être affinées au fil de plusieurs sessions de cartographie (c'est à dire au cours de différentes campagnes d'acquisition à des périodes et durées différentes). Enfin, lors de l'exploration sur différentes périodes et durées, le graphe initial construit est sujet à modifications. Des nœuds peuvent être ajoutés au graphe quand une nouvelle zone est découverte. La proposition de [Kretzschmar *et al.* 2010] permet de répondre à la prévention de l'explosion du nombre de nœuds dû à l'ajout de nœuds redondants, en élaguant le graphe de pose.

Bibliography

[Achanta *et al.* 2012] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua and S. Süsstrunk. *SLIC superpixels compared to State-of-the-Art Superpixel Methods* . IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI, vol. 34, no. 11, pages 2274–2281, 2012. (Cited on page 122.)

[Agarwal 2015] P. Agarwal. *Robust Graph-Based Localization and Mapping*. PhD thesis, University of Freiburg, 2015. (Cited on page 111.)

[AHS 2014] *Top Driverless Trucks in the Mining Industry Today Plus Future Concepts*. http://www.miningglobal.com/machinery/947/Top-Driverless-Trucks-in-the-M July 2014. (Cited on pages 1 and 9.)

[Aly *et al.* 2011] M. Aly, M. Munich and P.Perona. *Indexing in large scale image collections: Scaling properties and benchmark*. IEEE Computer Society, 2011. (Cited on page 32.)

[Andrienko *et al.* 2010] G. Andrienko, N. Andrienko, M. Mladenov, M. Mock and C. Poelitz. *Extracting Events from Spatial Time Series*. In 14t̂h Int. Conf. on Information Visualisation (IV), 2010. (Cited on page 104.)

[Atkinson & Shiffrin 1968] R.C Atkinson and R.M Shiffrin. The Psychology of Learning and Motivation. New York: Academic Press, 1968. (Cited on pages 29 and 31.)

[Bacca 2012] B. Bacca. *Appearance-based Mapping and Localization using Feature Stability Histrograms for Mobile Robot Navigation*. PhD thesis, Universitat de Girona, 2012. (Cited on pages 31, 133 and 138.)

[Baddeley 2003] A. Baddeley. *WORKING MEMORY: LOOKING BACK AND LOOKING FORWARD*. Nat Rev Neurosci, vol. 4, pages 829–839, 2003. (Cited on page 31.)

[Badino *et al.* 2011] H. Badino, D. Huber and T. Kanade. *Visual Topometric Localization*. In Intelligent Vehicles Symposium, Baden Baden, Germany, 2011. (Cited on page 27.)

[Bailey 2002] T. Bailey. *Mobile Robot Localisation and Mapping in Extensive Outdoor Environments*. PhD thesis, Australian Centre for Field Robotics, University of Sydney, 2002. (Cited on page 27.)

[Baker & Matthews 2001] S. Baker and I. Matthews. *Equivalence and efficiency of image alignment algorithms*. IEEE Intl. Conf. on Computer Vision, ICCV, vol. 1, no. 1090, 2001. (Cited on pages 20 and 75.)

[Baker & Matthews 2004] S. Baker and I. Matthews. *Lucas-Kanade 20 Years on: A Unifying Framework*. International Journal of Computer Vision, vol. 56, no. 3, pages 221–255, Feb 2004. (Cited on page 68.)

[Bay *et al.* 2006] H. Bay, T. Tuytelaars and L.V. Gool. *Surf: Speeded up robust features*. In European Conference on Computer Vision, ECCV, pages 404–417, 2006. (Cited on page 19.)

[Bebis 2012] G. Bebis. *Camera Calibration*. Rapport technique, Class notes for CS485/685, Department of Computer Science, University of Nevada, Spring 2012. (Cited on page 42.)

[Benhimane & Malis 2004] S. Benhimane and E. Malis. *Real-time image-based tracking of planes using efficient second-order minimization*. In IEEE Intl. Conf. Intelligent Robots and Systems, IROS, 2004. (Cited on page 68.)

[Besl & McKay 1992] P.J. Besl and N. McKay. *A method for registration of 3-D shapes*. IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI, vol. 14, no. 2, pages 239–256, 1992. (Cited on pages 80, 81 and 82.)

[Biber & Duckett 2009] P. Biber and T. Duckett. *Experimental Analysis of Sample-Based Maps for Long-Term SLAM*. International Journal of Robotics Research, vol. 28, no. 1, pages 20–33, Jan 2009. (Cited on page 31.)

[Blais & Levine 1995] G. Blais and M.D. Levine. *Registering Multiview Range Data to Create 3D Computer Objects*. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 17, no. 8, pages 820–824, 1995. (Cited on page 81.)

[Blanco 2010] José-Luis Blanco. *A tutorial on SE(3) transformation parameterizations and on-manifold optimization*. Rapport technique, University of Malaga, 2010. (Cited on pages 71 and 117.)

[Burt & Adelson] P.J Burt and E.H Adelson. *A multiresolution spline with application to image mosaics*. ACM Transactions on Graphics, vol. 2, pages 217–236. (Cited on page 76.)

[Chapoulie *et al.* 2011] A. Chapoulie, P. Rives and D. Filliat. *A spherical representation for efficient visual loop closing*. In IEEE Computer Vision Workshops, ICCV, 2011. (Cited on pages 26, 102, 133 and 137.)

[Chapoulie *et al.* 2012] A. Chapoulie, P. Rives and D. Filliat. *Topological segmentation of indoors/outdoors sequences of spherical views*. In International Conference on Robots and Systems, (IROS), 2012. (Cited on page 26.)

[Chapoulie *et al.* 2013] A. Chapoulie, P. Rives and D. Filliat. *Appearance-based segmentation of indoors/outdoors sequences of spherical views*. In International Conference on Robots and Systems, (IROS), 2013. (Cited on pages 23 and 26.)

[Chen & Medioni 1992] Y. Chen and G. Medioni. *Object modelling by registration of multiple range images*. Image and Vision Computing, vol. 10, no. 3, pages 145–155, 1992. (Cited on pages 80, 82 and 84.)

[Cheng *et al.* 2006] Y. Cheng, M.W. Maimone and L. Matthies. *Visual Odometry on the Mars Exploration Rovers*. IEEE Robotics and Automation Magazine, RAM, pages 54–62, 2006. (Cited on page 18.)

[Davison *et al.* 2007] A. Davison, I. Reid, N.D. Molton and O. Stasse. *MonoSLAM: Real-time single camera SLAM*. IEEE Trans. on Pattern Analysis and Machine Intelligence, (PAMI), pages 1052–1067, June 2007. (Cited on page 22.)

[Dayoub *et al.* 2011] F. Dayoub, G. Cielniak and T. Duckett. *Long- Term Experiments with and Adaptive Spherical View Representation for Navigation in Changing Environment*. Robotics and Autonomous Systems, vol. 59, no. 5, May 2011. (Cited on pages 29, 30, 32, 133 and 137.)

[Dayoub *et al.* 2013] F. Dayoub, T. Morris, B. Upcroft and P. Corke. *Vision-Only Autonomous Navigation Using Topometric Maps*. In International Conference on Robots and Systems, (IROS), 2013. (Cited on page 27.)

[Dellaert & Collins 1999] F. Dellaert and R. Collins. *Fast Image-based tracking by selective pixel integration*. In Proc. of the ICCV Workshop on frame-rate vision, September 1999. (Cited on page 74.)

[Drouilly *et al.* 2013] R. Drouilly, P. Rives and B. Morisset. *Fast Hybrid Relocation in Large Scale Metric-Topologic-Semantic Map*. In International Conference on Robots and Systems, (IROS), 2013. (Cited on page 27.)

[Dryanovski *et al.* 2013] I. Dryanovski, R.G. Valenti and J. Xiao. *Fast Visual Odometry and Mapping from RGB-D data*. In International Conference on Robotics and Automation. IEEE/RSJ, May 2013. (Cited on pages 79, 113 and 118.)

[Duda *et al.* 2001] R.O. Duda, P.E. Hart and D.G Stork. Pattern Classification, 2nd Ed. John Wiley and Sons, 2001. (Cited on page 120.)

[Durrant-Whyte & Bailey 2006] H. Durrant-Whyte and T. Bailey. *Simulataneous Localization and Mapping: Part I*. IEEE Robotics and Automation Magazine, pages 99–108, June 2006. (Cited on page 25.)

[Durrant-Whyte 1996] H. Durrant-Whyte. *An Autonomous Guided Vehicle for Cargo Handling Applications*. Intl. Journal of Robotics Research, (IJRR), vol. 15, no. 5, pages 407–440, October 1996. (Cited on pages 1 and 9.)

[Eade & Drummond 2007] Ethan Eade and Tom Drummond. *Monocular SLAM as a Graph of Coalesced Observations*. In Proc. 11th IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil, October 2007. (Cited on page 22.)

[Elfes 1989] A. Elfes. *Using occupancy grids for mobile robot perception and navigation*. Computer, vol. 22, no. 6, 1989. (Cited on page 23.)

[Endres *et al.* 2014] F. Endres, J. Hess, J. Sturm, D. Cremers and W. Burgard. *3D Mapping with an RGB-D Camera*. IEEE Trans. on Robotics, vol. 30, no. 1, 2014. (Cited on page 79.)

[Engel *et al.* 2014] J. Engel, T. Schöps and D. Cremers. *LSD-SLAM: Large-Scale Direct Monocular SLAM*. In European Conference on Computer Vision, 2014. (Cited on pages 133 and 137.)

[EPo] *The European Technology Platform on Smart Systems Integration*. http://www.smart-systems-integration.org/public. (Cited on pages 3 and 11.)

[Fairfield 2009] N. Fairfield. *Localization, Mapping and Planning in 3D Environments*. PhD thesis, Robotics Institute, Carnegie Mellon University, 2009. (Cited on page 26.)

[Fernández-Moral *et al.* 2014] E. Fernández-Moral, J. González-Jiménez, P. Rives and V. Arévalo. *Extrinsic calibration of a set of range cameras in 5 seconds without pattern*. In International Conference on Intelligent Robots and Systems. IEEE/RSJ, sep 2014. (Cited on pages 58 and 60.)

[Fischler & Bolles 1981] M.A. Fischler and R.C. Bolles. *RANSAC sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography*. Commun, ACM, vol. 24, no. 6, pages 381–395, 1981. (Cited on page 18.)

[Fitzgibon 2003] A. W. Fitzgibon. *Robust registration of 2D and 3D point sets*. Image and Vision Computing, vol. 21, pages 1145–115, 2003. (Cited on pages 19 and 63.)

[Fraundorfer & Scaramuzza 2012] F. Fraundorfer and D. Scaramuzza. *Visual Odometry: PartII: Matching, Robustness, Optimisation and Applications*. IEEE Robotics and Automation magazine, vol. 19, no. 2, pages 78–90, June 2012. (Cited on pages 19, 59, 63 and 74.)

[Fusiello *et al.* 2000] A. Fusiello, E. Trucco and A. Verri. *A compact algorithm for rectification of stereo pairs*. Machine Vision and Applications, Springer-Verlag, vol. 12, march 2000. (Cited on page 47.)

[Gca 2011] *Google Lobbies Nevada to Allow Self-Driving Cars*. http://www.nytimes.com/2011/05/11/science/11drive.html?_r=2&emc=eta1&, 10, May 2011. (Cited on pages 2 and 10.)

[Geiger *et al.* 2010] A. Geiger, M. Roser and R. Urtasun. *Efficient large-scale stereo matching*. In Asian Conference on Computer Vision, (ACCV), 2010. (Cited on pages 56 and 61.)

[Gokhool *et al.* 2014] T. Gokhool, M. Meilland, P. Rives and E. Fernández-Moral. *A Dense Map Building Approach from Spherical RGBD Images*. In International Conference on Computer Vision Theory and Applications,(VISAPP), Lisbon, Portugal, January 2014. (Cited on pages 7 and 14.)

[Gokhool *et al.* 2015] T. Gokhool, R. Martins, P. Rives and N. Despré. *A Compact Spherical RGBD Keyframe-based Representation*. In International Conference on Robotics and Automation,(ICRA), Washington, US, May 2015. (Cited on pages 7 and 15.)

[Grisetti *et al.* 2007] G. Grisetti, C. Stachniss, S. Grzonka and W. Burgard. *A tree parameterization for efficiently computing maximum likelihood maps using gradient descent*. In Proceedings of Robotics: Science and Systems, RSS, 2007. (Cited on page 111.)

[Haralick *et al.* 1989] R.M. Haralick, H. Joo, C-N. Lee, X. Zhuang, V.G. Vaidya and M.B. Kim. *Pose Estimation from Corresponding Point Data*. IEEE Trans. on Systems, Man and Cybernetics, vol. 19, no. 6, pages 1426–1446, December 1989. (Cited on pages 18, 80 and 81.)

[Harris & Stephens 1988] C. Harris and M. Stephens. *A Combined Corner and Edge Detector*. In 4th Alvey Vision Conference, 1988. (Cited on page 19.)

[Hartley & Zisserman 2003] R. Hartley and A. Zisserman. Multiple View Geometry in computer vision. Cambridge University Press, 2003. (Cited on page 66.)

[Harville *et al.* 1999] M. Harville, A. Rahimi, T. Darrell, G. Gordon and J. Woodfill. *3D Pose Tracking with Linear Depth and Brightness Constraints*. In IEEE Intl. Conf. on Computer Vision, ICCV, volume 1, pages 206–213, 1999. (Cited on pages 65 and 84.)

[Heckbert 1989] P.S. Heckbert. Fundamentals of texture mapping and image warping. Master's thesis, 1989. (Cited on page 51.)

[Henrichsen 2000] A. Henrichsen. 3-D reconstruction and camera calibration from 2-d images. Master's thesis, 2000. (Cited on page 37.)

[Henry *et al.* 2012] P. Henry, M. Krainin, E. Herbst, X. Ren and D. Fox. *RGB-D Mapping: Using Kinect-Style Depth Cameras for Dense 3D Modeling of Indoor Environments* . International Journal of Robotics Research, vol. 31, no. 5, pages 647–663, April 2012. (Cited on pages 22, 84, 133 and 137.)

[Herbst *et al.* 2011] E. Herbst, P. Henry, X. Ren and D. Fox. *Toward Object Discovery and Modeling via 3-D Scene Comparison*. In Intl. Conf. on Robotics and Automation, (ICRA), 2011. (Cited on pages 133 and 137.)

[Hirschmuller 2006] H. Hirschmuller. *Stereo processing by semi global block matching and mutual information*. IEEE Trans. on Pattern Analysis and Machine Intelligence, (PAMI), vol. 30, pages 328–341, 2006. (Cited on pages 56, 61 and 101.)

[Horn 1987] B. Horn. *Closed-form solution of absolute orientation using unit quaternions*. Journal of Optical Society, 1987. (Cited on pages 133 and 137.)

[Huber 1981] P.J. Huber. Robust Statistics. New York, Wiley, 1981. (Cited on page 74.)

[Johannsson *et al.* 2013] H. Johannsson, M. Kaess, M.F. Fallon and J.J. Leonard. *Temporally Scalable Visual SLAM using a Reduced Pose Graph*. In IEEE Intl. Conf. on Robotics and Automation, ICRA, Karlsruhe, Germany, May 2013. To appear. (Cited on page 33.)

[Johns & Yang 2014] E. Johns and G-Z. Yang. *Generative Methods for Long-Term Place Recognition in Dynamic Scenes*. International Journal of Computer Vision (IJCV), vol. 106, no. 3, pages 297–314, 2014. (Cited on pages 123, 133 and 137.)

[Kaess *et al.*] M. Kaess, A. Ranganathan and F. Dellaert. *iSAM: Incremental Smoothing and Mapping*. IEEE Transactions on Robotics, vol. 24, no. 6, pages 1365–1378. (Cited on page 33.)

[Kerl *et al.* 2013a] C. Kerl, J. Sturm and D. Cremers. *Dense Visual SLAM for RGB-D Cameras*. In Proc. of the Int. Conf. on Intelligent Robot Systems (IROS), Tokyo, Japan, 2013. (Cited on pages 79, 84, 86, 133 and 137.)

[Kerl *et al.* 2013b] C. Kerl, J. Sturm and D. Cremers. *Robust Odometry Estimation for RGB-D cameras*. In Proc. of the Int. Conf. on Robotics and Automation (ICRA), Karlsruhe, Germany, 2013. (Cited on pages 87 and 124.)

[Keys 1981] R. Keys. *Cubic convolution interpolation for digital image processing*. IEEE Transactions on Speech and Signal Processing, vol. 29, pages 1153–1160, 1981. (Cited on page 52.)

[Khoshelham & Elberink 2012] Kourosh Khoshelham and Sander O. Elberink. *Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications*. Sensors, vol. 12, no. 2, 2012. (Cited on pages 79, 107, 116 and 117.)

[Kim & Eustice 2013] A. Kim and R. Eustice. *Real-time visual SLAM for autonomous underwater hull inspection using visual saliency*. IEEE Transactions on Robotics, vol. 29, no. 3, pages 719–733, 2013. (Cited on page 86.)

[Klein & Murray 2007] G. Klein and D.W. Murray. *Parallel Tracking and Mapping for Small AR Workspaces*. In Proceedings of 6th IEEE Symp on mixed and Augmented Reality, Nara, Japan, November 2007. (Cited on page 22.)

[Klein & Murray 2008] G. Klein and D.W. Murray. *Improving the Agility of Keyframe-based SLAM*. In European Conference on Computer Vision, (ECCV), 2008. (Cited on page 22.)

[Konolige & Agrawal 2008] K. Konolige and M. Agrawal. *FrameSLAM: From Bundle Adjustment to Real-Time Visual Mapping*. IEEE Transaction on Robotics, vol. 24, no. 5, October 2008. (Cited on page 33.)

[Konolige & Bowman 2009] K. Konolige and J. Bowman. *Towards lifelong Visual Maps*. In International Conference on Intelligent Robots and Systems, October 2009. (Cited on pages 33, 86 and 123.)

[Konolige 2010] K. Konolige. *Sparse sparse bundle adjustment*. In Proc. of British Machine Vision Conference, BMVC, 2010. (Cited on page 111.)

[Kretzschmar *et al.* 2010] H. Kretzschmar, G. Grisetti and C. Stachniss. *Lifelong map learning for Graph-based SLAM in static environments*. Künstliche Intelligenz, KI, vol. 24, no. 3, pages 199–206, 2010. (Cited on pages 32, 86, 134 and 138.)

[Kümmerle *et al.* 2011] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige and W. Burgard. *g2o: A general framework for graph optimization*. In IEEE Intl. Conf. on Robotics and Automation, ICRA, 2011. (Cited on pages 102, 111, 133 and 137.)

[Lacroix *et al.* 1999] S. Lacroix, A. Mallet, R. Chatila and L. Gallo. *Rover Self Localization in Planetary-Like Environments*. In 5th Int. Symp. on Articial Intelligence, Robotics and Automation in Space, June 1999. (Cited on page 18.)

[Leutenegger *et al.* 2011] S. Leutenegger, M. Chli and R. Siegwart. *Brisk: Binary robust invariant scalable keypoints*. In IEEE Intl. Conf. on Computer Vision, ICCV, pages 2548–2555, 2011. (Cited on page 19.)

[Levoy *et al.* 2000] M. Levoy, K. Pulli, C. Curless, S. Rusinkiewicz, D. Koller, L.Pereira, M. Ginzton, S.E. Anderson, J. Davis, J. Shade and D. Fulk. *The digital Michelangelo Project: 3D scanning of large statues*. In Proc. of SIGGRAPH, 2000. (Cited on pages 24 and 80.)

[Lhuillier & Perriollat 2006] M. Lhuillier and M. Perriollat. *Uncertainty ellipsoids calculations for complex 3D reconstruction*. In IEEE International Conference on Robotics and Automation (ICRA), Orlando, May 2006. (Cited on page 21.)

[Llinas 2002] R.R. Llinas. *I of the Vortex: From Neurons to Self*. 2002. (Cited on page 31.)

[Lovegrove & Davison 2010] S. Lovegrove and A.J. Davison. *Real-Time Spherical Mosaicing using Whole Image Alignment*. In European Conf. on Computer Vision (ECCV), 2010. (Cited on page 52.)

[Low 2004] K-L. Low. *Linear Least-Squares Optimization for Point-to-Plane Surface Registration*. Rapport technique TR04-004, University of North Carolina at ChapelHill, 2004. (Cited on page 83.)

[Lowe 2003] D. Lowe. *Distinctive image features from scale-invariant keypoints*. Int. Journal of Computer Vision, IJCV, vol. 20, no. 2, pages 91–110, 2003. (Cited on page 19.)

[Lucas & Kanade 1981] B. Lucas and T. Kanade. *An iterative image registration technique with an application to stereo vision*. In International Joint Conference on Artificial Intelligence, pages 674–679, 1981. (Cited on pages 20, 64, 66 and 73.)

[Lui *et al.* 2012] W.L.D. Lui, T.J.J. Tang, T. Drummond and W.H. Li. *Robust Egomotion Estimation using ICP in Inverse Depth Coordinates*. In IEEE Intl. Conf. on Intelligent Robots and Systems, IROS, Villamoura, Portugal, 2012. (Cited on page 82.)

[Ma *et al.* 2004] Y. Ma, S. Soatto, J. Košecká and Shankar S Sastry. An invitation to 3-d vision. Springer, 2004. (Cited on pages 46, 49, 50, 52, 71 and 72.)

[Malis 2004] E. Malis. *Improving vision-based control using efficient second-order minimization techniques*. In IEEE Intl. Conf. Robotics and Automation, ICRA, volume 2, pages 1843–1848, 2004. (Cited on pages 20 and 68.)

[Martins *et al.* 2015] R. Martins, E. Fernández-Moral and P. Rives. *Dense accurate urban mapping from spherical RGB-D images*. In IEEE Intl. Conf. on Robots and Systems, IROS, 2015. submitted. (Cited on page 122.)

[Martins 2015] R. Martins. *RGBD Sphere Segmentation on Patches for Precise and Compact Scene Representation*. Rapport technique, INRIA, Sophia Antipolis, 2015. (Cited on page 101.)

[Matthies 1980] L. Matthies. *Dynamic stereo vision*. PhD thesis, Cargenie Mellon University, CMU, Pittsburgh, PA, 1980. (Cited on page 18.)

[McDonald *et al.* 2013] J.B. McDonald, M. Kaess, C. Cadena, J. Neira and J.J. Leonard. *Real-time 6-DOF Multi-session Visual SLAM over Large Scale Environments*. Journal of Robotics and Autonomous Systems, RAS, 2013. To appear. (Cited on page 34.)

[Mei 2007] C. Mei. *Laser-Augmented Omnidireactional Vision for 3D Localisation and Mapping*. PhD thesis, Ecole Nationale Superieure de mines de Paris, 2007. (Cited on page 38.)

[Meilland & Comport 2013a] M. Meilland and A. Comport. *On unifying keyframe and voxel based dense visual SLAM at large scales*. In IEEE Int. Conf. on Robots and Systems (IROS), Tokyo, Japan, October 2013. (Cited on pages 84 and 113.)

[Meilland & Comport 2013b] M. Meilland and A.I. Comport. *Simultaneous super-resolution tracking and mapping*. In IEEE Intl. Conf. on Robotics and Automation, ICRA, Karslruhe, Germany, May 2013. (Cited on pages 20, 63 and 90.)

[Meilland *et al.* 2010] M. Meilland, A.I. Comport and P. Rives. *A Spherical Robot-Centered Representation for Urban Navigation*. In IEEE Intl. Conf. on Intelligent Robots and Systems, IROS, Taiwan, October 18-22 2010. (Cited on pages 4, 12, 38, 53, 75, 114 and 124.)

[Meilland *et al.* 2011a] M. Meilland, A.I. Comport and P. Rives. *Dense visual mapping of large scale environments for real-time localisation*. In IEEE Intl. Conf. on Intelligent Robots and Systems, IROS, San Francisco, USA, 2011. (Cited on pages 4, 12, 38, 53, 69, 86, 87, 115 and 117.)

[Meilland *et al.* 2011b] M. Meilland, A.I Comport and P. Rives. *Real-time direct model-based tracking under large lighting variation*. British Machine Vision, 2011. (Cited on pages 4 and 12.)

[Meilland 2012] M. Meilland. *Cartographie RGB-D dense pour la localistation visuelle temps-réel et la navigation autonome*. PhD thesis, Ecole Nationale Superieure de mines de Paris, 2012. (Cited on page 87.)

[Moravec 1980] H. Moravec. *Obstacle avoidance and navigation in the real world by a seeing robot rover*. PhD thesis, Stanford University, 1980. (Cited on pages 17 and 18.)

[Morris *et al.* 2014a] T. Morris, F. Dayoub, P. Corke and B. Upcroft. *Simultaneous Localization and Planning on Multiple Map Hypotheses*. In IEEE Intl. Conf. on Robots and Systems, IROS, Chicago, USA, 2014. (Cited on page 30.)

[Morris *et al.* 2014b] T. Morris, F. Dayoub, P. Corke, G. Wyeth and B. Upcroft. *Multiple map hypotheses for planning and navigating in non-stationary environments*. In IEEE Intl. Conf. on Robotics and Automation, ICRA, HongKong, China, 2014. (Cited on page 30.)

[Morvan 2009] Y. Morvan. *Acquisition, Compression and Rendering of Depth and Texture for Multi-View video*. PhD thesis, Technical University of Eindhoven, 2009. (Cited on pages 37, 46 and 47.)

[Mouragnon *et al.* 2006a] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser and P. Sayd. *Monocular Based Vision SLAM for Mobile Robots* . In Proceedings of the IAPR International Conference on Pattern Recognition, Hong Kong, August 2006. (Cited on page 21.)

[Mouragnon *et al.* 2006b] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser and P. Sayd. *Real Time Localization and 3D Reconstruction* . In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New York, June 2006. (Cited on page 21.)

[MRG] *Oxford Mobile Robotics Group*. `http://www.http://mrg.robots.ox.ac.uk/`. (Cited on pages 2 and 10.)

[Murarka *et al.* 2006] A. Murarka, J. Modayil and B. Kuipers. *Building Local Safety Maps for a Wheelchair Robot using Vision Lasers*. In 3rd Canadian Coference on Computer and Robot Vision (CRV), 2006. (Cited on page 119.)

[Neugebauer 1997] P.J. Neugebauer. *Geometrical cloning of 3D objects via simultaneous registration of multiple range images*. In Proc. of the 1997 International Conference on Shape Modeling and Applications, 1997. (Cited on page 81.)

[Newcombe *et al.* 2011] R.A Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A.J. Davison, P. Kohli, J. Shotton, S. Hodges and A. Fitzgibbon. *KinectFusion: Real-Time Dense Surface Mapping and Tracking*. In IEEE Intl. Symp. on Mixed and Augmented Reality, ISMAR, October 2011. (Cited on pages 24, 81, 82, 83 and 84.)

[Newcombe 2012] R.A Newcombe. *Dense Visual SLAM*. PhD thesis, Imperial College London, December 2012. (Cited on pages 20 and 61.)

[Nguyen *et al.* 2012] C.V. Nguyen, S. Izadi and D. Lovell. *Modeling Kinect Sensor Noise for Improved Reconstruction and Tracking*. In IEEE Second Joint Conference in 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIM/3DPVT), 2012. (Cited on page 117.)

[Nistér *et al.* 2004] D. Nistér, O. Naroditsky and J. Bergen. *Visual Odometry*. IEEE Intl. Conf. on Computer Vision, ICCV, 2004. (Cited on page 18.)

[Nistér 2001] D. Nistér. *Frame decimation for structure and motion*. In in 2nd Workshop on Structure from Multiple Images of Large Environments, Springer Lecture Notes on Computer Science, volume 2018, pages 17–34, 2001. (Cited on page 21.)

[Nistér 2003] D. Nistér. *Preemptive RANSAC for Live Structure and Motion Estimation*. IEEE Intl. Conf. on Computer Vision, ICCV, pages 199–206, 2003. (Cited on page 19.)

[Olson 2008] E. Olson. *Robust and Efficient Robotic Mapping*. PhD thesis, Massachusetts Institute of Technology, 2008. (Cited on pages 25 and 33.)

[pan 2012] *Panoramic Photography*. `http://en.wikipedia.org/wiki/Panoramic_photography`, mar 2012. (Cited on page 38.)

[Park *et al.* 2012] J-H. Park, Y-D. Shin, J-H. Bae and M-H. Baeg. *Spatial Uncertainty Model for Visual Features Using Kinect Sensor*. Sensors, vol. 12, pages 8640–8662, 2012. (Cited on page 79.)

[Paul & Newman 2010] R. Paul and P. Newman. *FAB-MAP 3D: Topological Mapping with Spatial Visual Appearance*. In Intl. Conf. on Robotics and Automation, (ICRA), 2010. (Cited on pages 133 and 137.)

[Pizzoli *et al.* 2014] M. Pizzoli, C. Forster and D. Scaramuzza. *REMODE: Probabilistic Dense Reconstruction in Real Time*. In International Conference on Robotics and Automation,(ICRA), Hong Kong, China, May 2014. (Cited on page 20.)

[Pronobis & Jensfelt 2012] A. Pronobis and P. Jensfelt. *Large-scale Mapping and Reasoning with Heterogeneous Modalities*. In International Conference on Robotics and Automation,(ICRA), Minnesota, USA, 2012. (Cited on page 28.)

[Rahimi *et al.* 2001] A. Rahimi, L.P. Morency and T. Darrell. *Reducing Drift in Parametric Motion Tracking*. In IEEE Intl. Conf. on Computer Vision, ICCV, 2001. (Cited on page 84.)

[Rives *et al.* 2014] P. Rives, R. Drouilly and T. Gokhool. *Représentation orientée navigation d'environnements à grande échelle*. In Reconnaissance de Formes et Intelligence Artificielle, RFIA 2014, France, June 2014. (Cited on pages 7 and 14.)

[Rosten & Drummond 2006] E. Rosten and T. Drummond. *Machine learning for high speed corner detection*. In European Conference on Computer Vision, ECCV, 2006. (Cited on page 19.)

[Royer *et al.* 2007] E. Royer, M. Lhuillier, Dhome M. and Lavest J-M. *Monocular Vision for Mobile Robot Localization and Autonomous Navigation*. International Journal of Computer Vision, vol. 74, no. 3, pages 237 – 260, January 2007. (Cited on pages 21 and 86.)

[Royer 2006] E. Royer. *Cartographie 3D et Localisation par vision monoculaire pour la navigation autonome d'un robot mobile*. PhD thesis, Université Blaise Pascal-Clermont II, 2006. (Cited on page 21.)

[Rublee *et al.* 2011] E. Rublee, V. Rabaud, K. Konolige and G. Bradski. *Orb: An efficient alternative to sift or surf (pdf)*. In IEEE Intl. Conf. on Computer Vision, ICCV, pages 2564–2571, 2011. (Cited on page 19.)

[Rusinkiewicz & Levoy 2001] S. Rusinkiewicz and M. Levoy. *Efficient Variants of the ICP Algorithm*. In Proc. of IEEE 3rd International Conference on 3D- Digital Imaging and Modeling, Quebec, CA, 2001. (Cited on pages 81 and 83.)

[Rusinkiewicz *et al.* 2002] S. Rusinkiewicz, O. Hall-Holt and M. Levoy. *Real-Time 3D Model Acquisition*. In Proc. of SIGGRAPH, 2002. (Cited on page 82.)

[Salas-Moreno *et al.* 2013] R.F. Salas-Moreno, R.A. Newcombe, H. Strasdat and P.H.J. Kelly. *SLAM++: Simultaneous Localisation and Mapping at the level of Objects*. In Computer Vision and Pattern Recognition,(CVPR), 2013. (Cited on page 28.)

[S.Baker & Matthews 2001] S.Baker and I. Matthews. *Equivalence and Efficiency of Image Alignment Algorithms*. In Proceedings of IEEE Computer Society on Computer Vision and Pattern Recognition-CVPR, pages 1090–1097, 2001. (Cited on page 26.)

[Scaramuzza & Fraundorfer 2011] D. Scaramuzza and F. Fraundorfer. *Visual Odometry: PartI: The First 30 Years and Fundamentals*. IEEE Robotics and Automation magazine, vol. 18, pages 80–92, December 2011. (Cited on pages 18, 79 and 80.)

[Schöps *et al.* 2014] T. Schöps, J. Engel and D. Cremers. *Semi-Dense Visual Odometry for AR on a Smartphone*. In IEEE Intl. Symp. on Mixed and Augmented Reality, ISMAR, Munich, Germany, 2014. (Cited on page 78.)

[Segal *et al.* 2009] A.V. Segal, D. Haehnel and S. Thrun. *Generalized-ICP*. In Robotics: Science and Systems, 2009. (Cited on page 82.)

[Steinbrücker *et al.* 2013] F. Steinbrücker, C. Kerl, J. Sturm and D. Cremers. *Large-Scale Multi-Resolution Surface Reconstruction from RGB-D Sequences*. In International Conference on Computer Vision, ICCV. IEEE/RSJ, 2013. (Cited on pages 24 and 81.)

[Stephen *et al.* 2002] S. Stephen, D. Lowe and J. Little. *Mobile Robot Localization and Mapping with Uncertainty using Scale-Invariant Visual Landmarks*. International Journal of Robotics Research (IJRR), vol. 21, no. 8, pages 735–758, August 2002. (Cited on page 113.)

[Strasdat *et al.* 2010] H. Strasdat, J.M.M Montiel and A.J Davison. *Scale Drift-Aware Large Scale Monocular SLAM*. In Robots Science and Systems, RSS, 2010. (Cited on page 86.)

[Strasdat 2012] H. Strasdat. *Local Accuracy and Global Consistency for Efficient Visual SLAM*. PhD thesis, Department of Computing, Imperial College London, October 2012. (Cited on page 22.)

[Sturm *et al.* 2012] J. Sturm, N. Engelhard, F. Endres, W. Burgard and D. Cremers. *A Benchmark for the Evaluation of RGB-D SLAM systems*. In IEEE/RSJ Intl. Conf. Intelligent Robots and Systems, IROS, Vilamoural, Portugal, 2012. (Cited on pages 89 and 90.)

[Sung 2008] E. Sung. *3-D Computer vision*. Rapport technique, Class notes for EE6222, Department of EEE, Nanyang Technological University, august 2008. (Cited on page 42.)

[Szeliski 2006] R. Szeliski. *Image alignment and stitching: a tutorial*. Foundations and Trends in Computer Graphics and Vision, vol. 2, no. 1, pages 1–104, January 2006. (Cited on page 53.)

[Thrun *et al.* 2005] S. Thrun, W. Burgard and D. Fox. Probabilistic Robotics. Cambridge: MIT Press, 2005. (Cited on page 79.)

[Tomasi & Shi 1994] C. Tomasi and J. Shi. *A Combined Corner and Edge Detector*. In IEEE Intl. Conf. on Computer Vision Computer Vision and Pattern Recognition, CVPR, pages 593–600, 1994. (Cited on page 19.)

[Torr *et al.*] P. Torr, A. Fitzgibbon and A. Zisserman. *The problem of degeneracy in structure and motion recovery from uncalibrated image sequences*. International Journal of Computer Vision, vol. 32, no. 1, pages 27–44. (Cited on page 21.)

[Triggs *et al.* 2000] B. Triggs, P.I. McLauchlan, H. Hartley and A.W Fitzgibbon. *Bundle Adjustment- A Modern Synthesis*. in W. Triggs, A. Zisserman, R. Szeliski (Eds) Vision Algorithms: Theory and Practice, in LNCS, vol. 1883, pages 298–372, 2000. (Cited on page 21.)

[Twinanda *et al.* 2013] Andru Putra Twinanda, Maxime Meilland, Désiré Sidibé and Andrew I. Comport. *On Keyframe Positioning for Pose Graphs Applied to Visual SLAM*. In 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, 5th Workshop on Planning, Perception and Navigation for Intelligent Vehicles, November 3rd 2013. (Cited on page 87.)

[Tykkälä *et al.* 2011] T. Tykkälä, C. Audras and A.I. Comport. *Direct Iterative Closest Point for Real-time Visual Odometry*. In IEEE Intl. Conf. on Computer Vision Workshops, ICCV, November 2011. (Cited on page 84.)

[Tykkälä *et al.* 2013] T. Tykkälä, H. Hartikainen, A.I. Comport and J.-K. Kämäräinen. *RGB-D Tracking and Reconstruction for TV Broadcasts*. In Int. Conf. on Computer Vision Theory and Applications (VISAPP), Barcelona, Spain, 2013. (Cited on page 84.)

[VCh] *Automated Valet Parking and Chargin for e-Mobility*. `http://www.v-charge.eu/`. (Cited on pages 3 and 11.)

[Vedula *et al.* 2005] S. Vedula, S. Baker, P. Rander, R. Collins and T. Kanade. *Three-Dimensional Scene Flow*. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 27, no. 3, pages 475–480, March 2005. (Cited on page 65.)

[Wang *et al.* 2006] Q. Wang, W. Zhang and X. Tang. *Real Time Bayesian 3-D Pose Tracking*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no. 12, pages 1533–1541, December 2006. (Cited on pages 84 and 86.)

[Wurm *et al.* 2010] K.M. Wurm, A. Hornung, M. Bennewitz, C. Stachniss and W. Burgard. *Octomap: A Probabilistic, Flexible and Compact 3D Map Representation for Robotic Systems*. In International Conference on Robotics and Automation,(ICRA), 2010. (Cited on page 24.)

[Zhang 1994] Z. Zhang. *Iterative Point Matching for Registration of Free-Form Curves*. International Journal of Computer Vision (IJCV), vol. 13, no. 2, pages 119–152, 1994. (Cited on pages 81 and 82.)

[Zhang 2000] Z. Zhang. *A Flexible New Technique for Camera Calibration*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 11, november 2000. (Cited on pages 42 and 59.)

[Zhao *et al.* 2005] W. Zhao, D. Nistér and S. Hsu. *Alignment of Continuous Video onto 3D Point Clouds*. IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, 2005. (Cited on pages 63, 79, 82 and 123.)

Cartographie dense basée sur une représentation compacte RGB-D dédiée à la navigation autonome

Résumé:

Dans ce travail, nous proposons une représentation efficace de l'environnement adaptée à la problématique de la navigation autonome. Cette représentation topométrique est constituée d'un graphe de sphères de vision augmentées d'informations de profondeur. Localement la sphère de vision augmentée constitue une représentation egocentrée complète de l'environnement proche. Le graphe de sphères permet de couvrir un environnement de grande taille et d'en assurer la représentation. Les "poses" à 6 degrés de liberté calculées entre sphères sont facilement exploitables par des tâches de navigation en temps réel. Dans cette thèse, les problématiques suivantes ont été considérées:

- Comment intégrer des informations géométriques et photométriques dans une approche d'odométrie visuelle robuste
- Comment déterminer le nombre et le placement des sphères augmentées pour représenter un environnement de façon complète
- Comment modéliser les incertitudes pour fusionner les observations dans le but d'augmenter la précision de la représentation
- Comment utiliser des cartes de saillances pour augmenter la précision et la stabilité du processus d'odométrie visuelle

Mots-clés: Capteurs RGB-D, SLAM visuel, odométrie visuelle, carte topométrique, image clés, représentation dense, association probabiliste, fusion de données.

A Compact RGB-D Map Representation dedicated to Autonomous Navigation

Abstract:

Our aim is concentrated around building ego-centric topometric maps represented as a graph of keyframe nodes which can be efficiently used by autonomous agents. The keyframe nodes which combines a spherical image and a depth map (augmented visual sphere) synthesises information collected in a local area of space by an embedded acquisition system. The representation of the global environment consists of a collection of augmented visual spheres that provide the necessary coverage of an operational area. A "pose" graph that links these spheres together in six degrees of freedom, also defines the domain potentially exploitable for navigation tasks in real time. As part of this research, an approach to map-based representation has been proposed by considering the following issues:

- How to robustly apply visual odometry by making the most of both photometric and geometric information available from our augmented spherical database
- How to determine the quantity and optimal placement of these augmented spheres to cover an environment completely
- How to model sensor uncertainties and update the dense infomation of the augmented spheres
- How to compactly represent the information contained in the augmented sphere to ensure robustness, accuracy and stability along an explored trajectory by making use of saliency maps

Keywords: RGB-D sensors, Visual SLAM, Visual odometry, topometric map, keyframe based map, dense mapping, probabilistic data association, sensor fusion